Working Paper No. 143

PROCESSING AND TRANSMITTING INFORMATION, GIVEN A PAY-OFF FUNCTION

by

HENRI MICHEL PHAM-HUU-TRI

December, 1968

OPERATIONS RESEARCH DIVISION
WESTERN MANAGEMENT SCIENCE INSTITUTE
University of California, Los Angeles

UNIVERSITY OF CALIFORNIA

Los Angeles

Processing and Transmitting Information, Given a Pay-Off Function

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy in Mathematics

by

Henri Michel Pham-Huu-Tri

Committee in charge:

Professor Jacob Marschak, Chairman

Professor A. V. Balakrishnan

Professor Jack W. Carlyle

Professor David Cantor

Professor Thomas S. Ferguson

1968

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

I would like to express my great admiration and my deep gratitude to Professor J. Marschak, not only for his guidance in my work, but for his whole life-style, which has been and will be an inspiration to me.

I wish, also, to thank Douglas Adkins for reading my manuscript and Elaine for her wonderful typing.

This work was supported by the Western Management Science Institute principally under Grants 1191 and 2041 from the National Science Foundation, and partially under a Ford Foundation Grant and Office of Naval Research Contract 233(75).

ABSTRACT OF THE DISSERTATION

Processing and Transmitting Information,

Given a Pay-Off Function

by

Henri Michel Pham-Huu-Tri

Doctor of Philosophy in Mathematics

University of California, Los Angeles, 1968

Professor Jacob Marschak, Chairman

An information system is defined as a chain of information services, encoding (processing)...transmitting...decoding (deciding). Each service is a transformer represented, in general, by a stochastic matrix and a cost function. The inputs of "encoding" are the pay-off-relevant events. Actions are the output of decoding, actions and events determine the pay-off. The utility of the services to the user is a function of the pay-off and of the different costs. Efficiently choosing an information system is by definition choosing an information system which maximizes the expected utility.

Communication engineers restricted themselves to information systems with fixed transmitting (channel) and identically zero cost functions. Moreover, they equated the user's utility function with his pay-off function. They handled the problem in the following way:

1. choose first encoding with respect to the source of events and the pay-off function only, 2. choose second encoding and decoding with respect to transmitting only. Encoding is the composition of first and second encoding. However, their approach was inefficient; 1. They neglected the pay-off function in the choice of second encoding and decoding, 2. they arbitrarily broke the original problem into two independent, more accessible, problems.

We also restricted ourselves to information systems with fixed transmitting and zero cost functions and users' utility functions identical to their pay-off functions. But our approach is more efficient because we treated the problem of choosing encoding and decoding, given a source of events, a pay-off function and a channel, as a whole. The bounds we obtained should, therefore, be better, at least in all cases where the pay-off function has a wide range of values. We did, however, treat the non-restricted problem with certain properties of the source, the channel and the utility function assumed.

SECTION 1

1.1. Introduction

The Economic Theory of Information is concerned with the efficient choice of Information services. J. Marschak (Efficient choice of Information Services, 1968. Conference for Research on Management Information Systems) distinguishes the following sequence of services in that order: Inquiring, communicating and deciding. Communicating is itself a sequence of Encoding, Transmitting and Decoding. Another component of the sequence, called Storing, which can be intermediate between any two consecutive services, will be disregarded in this work, together with Inquiring, which is the same as assuming that they are both costless and perfect. Moreover, Decoding and Deciding will be reduced, without loss of generality to a single operation: decoding into action. Our simplified chain of services, or information system, will then consist of only three links: Encoding, Transmitting, Decoding.

More precisely, see diagram 1.1, there will be a source, S, of events (or messages, since inquiring is assumed to be an identity operation) generating the random variable e from a finite set E with the distribution $P(\cdot)$. There will be discrete, memoryless channels denoted $(X, P(y|x), Y)$ or simply $\{P(y|x)\}$ with finite

input and output alphabets X and Y respectively. The Transmission $T(\cdot) : X \to Y$, $T(\cdot)$ is a random function; the Encoding will be denoted $\psi_1(\cdot) : E \to X$; the Decoding $\psi_2(\cdot) : Y \to A$ where A is the finite set of feasible actions {a}.

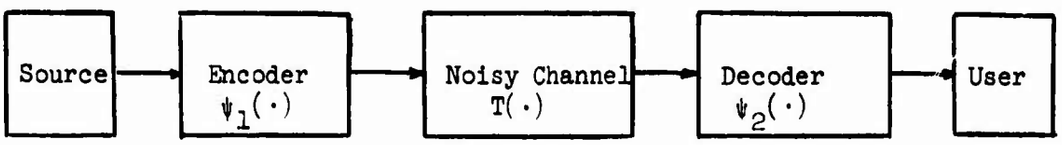


Diagram 1.1

One of the criteria, that will be considered in the choice of services is the benefit to the user, a function $\omega(\cdot)$ of e and a, called pay-off function: $\omega(\cdot,\cdot) : E \times A \to \text{Reals}$. The others being the costs of the different operations. If costs would not depend on the chosen information system, the user would, by definition, prefer the system yielding the highest expected Pay-off:

$$E\{\omega(e.a)\} = \sum_{e,a} P(e)\ \text{Prob}_{\psi_1,T,\psi_2}(a|e)\omega(e,a).$$

The subscript is here to recall that the probability of action a, given event e, is a function of Encoding, Transmitting, and Decoding.

Now if the costs are introduced: $k_{\psi_1}(e)$, cost of Encoding e, $k_T(x)$, cost of Transmitting x; $k_{\psi_2}(y)$, cost of decoding y, the user would try to maximize the expected value of a certain function $U(\omega(e,a),k_{\psi_1}(e),k_T(x),k_{\psi_2}(y))$, by definition his utility function.

Not much can be said about $U(\ ,\ ,\ ,\ )$ besides the fact that it is increasing in $\omega(e,a)$ and decreasing in $k_{\Psi_1}(e)$, $k_T(x)$, $k_{\Psi_2}(y)$. Moreover, the costs themselves are not well known, especially the costs of Encoding and Decoding, depending upon this complexity. One has to resort to using arbitrary elementary (often linear) function to represent $U(\cdot)$, $k_{\Psi_1}(\cdot)$, $k_T(\cdot)$, $k_{\Psi_2}(\cdot)$ more or less realistically.

We are not ready at this point to approach the general problem except for a special case: binary symmetric memoryless source, finary symmetric pay-off function, binary symmetric memoryless channel, $U(\cdot)$ linear in $k_{\Psi_1}(\cdot)$, $k_T(\cdot)$, $k_{\Psi_2}(\cdot)$. In the rest of this work the transmission system costs will be assumed constant and the choice will be restricted to Transmission Systems with a fixed Channel. In other words, attention will be devoted to the following problem, a preliminary one: Find Encoding and Decoding procedures that would maximize the expected pay-off function $\omega(\cdot,\cdot)$. In doing so, we will get some insight into the original problem and some partial answers to it.

1.2. Pure Communication of Information vs. Communication of Information, Given a Pay-off Function

What is usually called Information Theory is essentially a theory of pure communication. It was principally started by C. E. Shannon in 1948 in his, "A Mathematical Theory of Communication".

Let e be a random variable generated by a source S, taking on a finite number of values: $1,\cdots,g,\cdots,G$ with probabilities

P(1),···,P(G). The uncertainty associated with e was quite arbitrarily defined to be the quantity:

$$H(e) = H(P(1),\cdots,P(G)) = -\sum_{g=1}^{G} P(g) \log P(g)$$

where - log (P(g)) was interpreted as the uncertainty associated with the event (e = g) or the uncertainty removed (or information conveyed) by revealing that e has taken on the value g. H(e) is also called Entropy or Information rate of S.

This measure depends only on the probability distribution of the messages. In particular, two messages with the same probability have their information characterized by the same number although they are not necessarily equally valuable to the user, for he evaluates the economic value of a message by the maximum profit he can make by using it. The value of a Source of Information, as far as the user is concerned, is measured by the maximum expected Pay-off it can bring him.

Shannon's further analysis of communication systems relies greatly on his measure of Information. In his 1948 model (diagram 1.2) a randomly produced message generated by a Source is encoded into a signal belonging to a specified set, called vocabulary. The encoded message is transmitted through a noisy channel, whose output is decoded. The objective is to select a vocabulary such that the probability of correctly identifying the input signal is as large as possible.

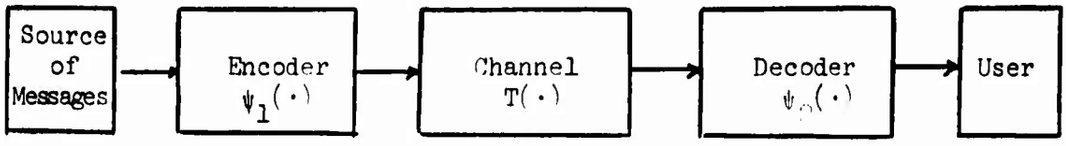


Diagram 1.2

Diagrams 1.1 and 1.2 are identical, but our objective is somewhat different: make the expected Pay-off as large as possible. However, they are not completely irreconcilable, as shown first by Shannon himself in his 1959 paper, "Coding theorems for a discrete source with a fidelity criterion". Besides, it is intuitively obvious that there should be some correlation between the probability of correct transmission and the optimal expected pay-off.

In his 1959 model, Shannon added a new component in his Communication System between the source and the Encoder and also a distortion function $d(e,a)$. $d(e,a)$ is the "cost" of taking action a when the message is e. In Economic terminology it is the loss, of not taking an optimal action.

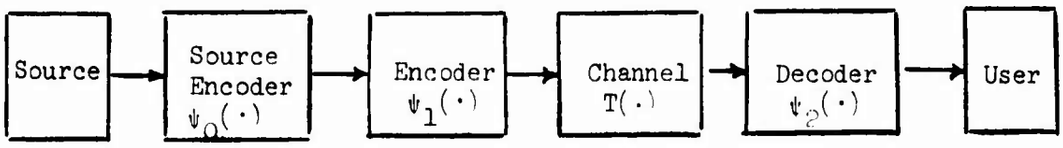


Diagram 1.3

This new operator $(\psi_0 : E \rightarrow A)$, mapped the messages e into a specified set of actions in such a way as to decrease the rate of information to be transmitted to a level acceptable to the channel, but resulting in some loss in pay-off. The actions were then transmitted with as small a probability of error as possible.

Later on, several authors, including Yudkin, Goblick, and Jelinek, improved the source-encoding procedure.

Let us point out that:

I. It is intuitively clear that their approach is inefficient because:

1) Double encoding $(\psi_0(\cdot), \psi_1(\cdot))$ is not justifiable although more accessible to mathematical study. Moreover, $\psi_0(\cdot)$ maps events directly onto actions. Thus, if an action maximizes the pay-off, given two different events, these two events will be encoded in the same message. This message will specify that particular action. Yet an error in transmitting that message will result in specifying a non-optimal action and thus may cause a much greater loss in the case of one event than in the case of the other. Two events e and e' equivalent with respect to optimal action, are, in general, not equivalent with respect to the values of $d(e,a)$, $d(e',a)$ for varying a. Thus $\psi_0(\cdot)$ would replace the set of "pay-off relevant events" by a generally coarser set of "action relevant events" and this diminishes the maximum expected pay-off (Reference I).

2) $\psi_1(\cdot)$ and $\psi_2(\cdot)$ are chosen with no account taken of the differences in losses due to having one rather than another

communication error.

3) They handled the communication problem in the following way: on the one hand, choose $\psi_0(\cdot)$ for the given source and loss function (in Shannon's terminology, the distortion function), $d(\cdot,\cdot)$, only, on the other hand, choose $\psi_1(\cdot)$ and $\psi_2(\cdot)$, given the channel only. However, breaking the communication problem into these two independent problems is not efficient in most cases. $\psi_1(\cdot)$ and $\psi_2(\cdot)$ should simultaneously be chosen given: S, $T(\cdot)$ and $d(\cdot,\cdot)$.

II. No explicit solutions are ever displayed, but only their existence is proved.

III. Only code words of fixed length are considered, although simple examples show that variable length encoders are often more efficient.

IV. The usual analysis is confined to long blocks of events and long code words which indeed tend to yield perfect results.

This last point is economically quite crucial. In practice, it is often impossible, or would result in great losses, to wait for a large number of messages to pile up before one starts to communicate them. In this sense, the information they carry might become obsolescent from the users point of view. A great deal of work is left to be done in this area.

We have made no progress with respect to II, III and IV. But we have given a special emphasis on non-asymptotic results, so that a user who can afford to wait for up to N messages to accumulate might

have some indication about how well he can do and about how to do it.

We have focussed our attention on I. In an effort to tie the given Source, Channel and Pay-off Function together, we have considered a deterministic correspondence (to be optimized) between channel input alphabet and the set of feasible actions. We are, therefore, able to ascribe a value (or loss) to each error, and thus to estimate the loss due to single-step (source and channel) encoding and also to increase the precision of the estimation of the loss due to transmission. In our procedure, both encoding and transmission aim to maximize the expected pay-off. In Shannon-Pilc-Jelinek's, only the encoding aimed to maximize the expected pay-off, while the transmission aimed to maximize the probability of correct transmission. Our upper-bound on the loss due to communication is therefore better in all cases where the loss function has a wide range of values.

1.3. Summary

In Section 2.1, a brief survey of the main concepts of Information Theory is made with an emphasis on a notion of special interest to us, the Rate-Distortion Function, introduced by Shannon in 1959, which we will call Rate-Loss Function. In section 2.2, we introduce our notation and definitions, set the relationship between Pay-off and Loss Functions, respectively $\omega(e,a)$ and $d(e,a)$, and describe our scheme. A Processing (Source Encoding) loss Function $\delta_1(e,x)$ and a Transmission Loss Function $\delta_2(x,x')$ are derived in such a way that

$$E_{e,a}\{d(e,a)|\psi_1(\cdot),T(\cdot),\psi_2(\cdot)\} \leq E_{e,x}\{\delta_1(e,x)|\psi_1(\cdot)\}$$

$$+ E_{xx'}\{\delta_2(x,x')|T(\cdot),\psi_2(\cdot)\}$$

It is convenient to consider the loss matrices associated with $d(\cdot,\cdot)$, $\delta_1(\cdot,\cdot)$ and $\delta_2(\cdot,\cdot)$. $[d(e,a)]$ (respectively $[\delta_1(e,x)]$, $[\delta_2(x,x)]$) is the matrix with $d(e,a)$ as entry in the e^{th} row and the a^{th} column.

In Section 3, we give a lower bound to the average loss one should expect with a channel of capacity C. Theorem 1 states that: for a constant, memoryless source S with a finite loss function $d(\cdot,\cdot)$, and a discrete, memoryless channel of capacity C, there exists no encoding and decoding procedure that yields an expected loss smaller than $R^{-1}(C)$, where $R^{-1}(\cdot)$ is the inverse function of $R(\cdot)$, the Rate-Loss Function, defined by Shannon. Corollary 1 states that: for a constant, memoryless source S with a finite loss function $d(\cdot,\cdot)$, there exists no source encoding procedure that yields a processing loss less than $R^{-1}(H(x))$ if $H(x)$ is the entropy of the channel input letters in the vocabulary.

Section 4 is devoted to Encoding. In 4.1 the source encoding (or Processing) procedure originated by Shannon and improved by Yudkin and Jelinek is described in detail and it is shown that there are encoding functions $\psi_1(\cdot)$ which yield an average Processing loss as close as we please to the lower bound of Corollary I. Theorem II is a converse

of Corollary I. In 4.2 a transmission loss function $\delta_2(\cdot,\cdot) : X \times X \to$ Reals, derived from the encoding Procedure, overbounds the loss when channel input x is sent over the channel and recovered as x'.

In Section 5 we prove a transmission loss theorem. Theorem III says roughly that it is possible to select vocabularies $u = \{u_1,\cdots,u_m,\cdots,u_M\}$, $M = e^{n\psi}$, of code words of length n, $u_m = (x_{m1},\cdots,x_{mn})$, and decoding functions $\psi_2(\cdot)$ that yield, on the average, a transmission loss as low as desired, provided $\psi < C$, the channel capacity.

In Section 6, Processing and Transmission are linked together to give Theorem IV and IV'. Theorems II, III, IV and IV' give in fact upper bounds to the various expected losses. Theorem IV states, in short, that there are codes $(\psi_1(\cdot),\psi_2(\cdot))$ that yield, on the average, a loss, due to communication, as close as desired to the lower bound in Theorem I, if Source, loss function and Channel are matched in a certain way. Theorem IV' is a variant of Theorem IV for limited length message blocks.

In Section 7 we treat the general problem stated in the introduction and give a tentative approach to the special case where the following additional assumptions obtain: binary, uniform source, binary symmetric loss function, binary symmetric channel, linear utility and cost functions.

SECTION 2

2.1. Basic Concepts of Information Theory

2.1.1. A discrete channel, denoted by $(X, p(y|x), Y)$ or $\{p(y|x)\}$ consists of two finite sets, X and Y, and a non-negative function $p(y|x)$, defined for all pairs (x,y), $x \in X$, $y \in Y$ such that $\Sigma\, p(y|x) = 1$ for all x's. X and Y are called input and output sets of the channel and $p(y|x)$ is the conditional probability to receive y when x is transmitted.

It is standard practice to consider the transmission of a sequence of symbols, each symbol belonging to the input set X. For any positive integer n and any set, for example X, we denote by X^n the set of n-tuples $(x_1,\cdots,x_n) = x^n$ with each $x_k \in X$. If a sequence $x^n = (x_1,\cdots,x_n)$ is applied at the input of the channel, then a sequence $y^n = (y_1,\cdots,y_n) \in Y^n$ is received at the output with a conditional probability $p(y_1,\cdots,y_n|x_1,\cdots,x_n)$ which has yet to be specified for all $x_1 \cdots x_n$ and all n. We will restrict our attention to discrete channels without memory. For such channels, successive operations are independent.

A discrete channel $(X, p(y|x), Y)$ is said to be memoryless if

$$p(y^n|x^n) = \prod_{k=1}^{n} p(y_k|x_k)$$

for all $y^n \in Y^n$ and all $x^n \in X^n$ and all $n \in \{1,2,\ldots\}$.

Thus a discrete, memoryless channel $(X,p(y|x),Y)$ is characterized by a matrix with row set X and column set Y, whose entries are $p(y|x)$. This matrix is called the channel matrix. In this work a channel will always be a discrete memoryless channel.

Let M and n be positive integers, and $0 \leq \lambda < 1$. A code of length M, word length n, and probability of error $\leq \lambda$, denoted (n,M,λ), for a discrete memoryless channel $(X,p(y|x),Y)$ consists of a sequence of M distinct elements of X^n, $\{u_1,\cdots,u_M\}$, and a sequence of M disjoint subsets of Y^n; $D_1,\cdots,D_M$; such that

$$P(D_m|u_m) = \sum_{y^n \in D_m} P(y^n|u_m) \geq 1 - \lambda \quad \text{for} \quad m = 1,\cdots,M,$$

$$P(y^n|u_m) = \prod_{k=1} P(y_k|x_k).$$

$\{u_1,\cdots,u_M\}$ is called a vocabulary of input messages or codewords and D_m is called a decoding set for u_m.

Practically one uses a code as follows. A message u_m is selected arbitrarily and tranmitted over the channel. The letter sequence y^n is received with probability $p(y^n|u_m)$. If $y^n \in D_t$ the receiver concludes that u_t was sent. The probability that any message u_m will be tranmitted so as to be decoded incorrectly is $\leq \lambda$.

A real number $\psi > 0$ is called an attainable transmission rate for a channel $p(y|x)$ if there exists a sequence of codes (n,M_n,λ_n) for $p(y|x)$ with $M_n \geq e^{n\psi}$ and $\lambda_n \to 0$. The transmission capacity ψ^* of a discrete memoryless channel is defined to be the

supremum of the set of its attainable rates. We may give the following interpretation to the transmission capacity ψ^*. If $0 < \psi < \psi^*$ then one can transmit any of ψ e-ary symbols per transmission period over the channel with an arbitrarily small probability of error by making the word length n large enough.

If $(X, p(y|x), Y)$ is a discrete memoryless channel and $q(x)$ is a given probability distribution on X, then we let $p(x,y) = p(y|x)q(x)$ and $r(y) = \sum_x p(x,y)$. We define

$$H(y) = -\sum_x r(y) \log r(y)$$

$$H(y|x) = -\sum_x q(x) \sum_y p(y|x) \log p(y|x)$$

where all logarithms are base e. Let

$$\begin{aligned} I(q) &= H(y) - H(y|x) \\ &= H(x) - H(x|y) \\ &= \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{q(x)r(y)} \end{aligned}$$

In information theory the quantity $I(q)$, which depends on the input distribution $q(\cdot)$ is interpreted as the average amount of information, per transmission, received through the channel. The maximum amount of information received through the channel is called channel capacity. It is defined as the maximum over $q(\cdot)$ of $I(q)$.

$$C = \max_{q} \sum_{x} \sum_{y} p(x,y) \log \frac{p(x,y)}{q(x)r(y)} = \sum_{x} \sum_{y} p(x,y) \log \frac{p(y|x)}{r(y)}$$

Where the max is taken over all distributions $q(\cdot)$ on X.

The fundamental Theorem of Information Theory which was first proved by Shannon states that, for any discrete memoryless channel, $\psi^* = C$.

2.1.2. Pay-off function and loss function. The pay-off function $\omega(\cdot,\cdot) : E \times A \to \text{Reals}$ gives the benefit associated with event e and action a. For any event e, there exists at least one optimal action $a(e)$ such that:

$$\omega(e,a(e)) \geq \omega(e,a) \ \forall \ a.$$

The loss function associated with $\omega(\cdot,\cdot)$ is defined on the same domain by the relation

$$d(e,a) \stackrel{\Delta}{=} \omega(e,a(e)) - \omega(e,a).$$

This function is what is called regret function in Decision Theory. We used the letter d because it plays exactly the same role, as far as processing and transmission of information go, as Shannon's "single letter distortion measure". We want to communicate information so as to maximize the expected-pay-off. It is actually the same to

communicate information so as to minimize the expected loss, or distortion.

2.1.3. The Rate-Loss Function (Rate-Distortion Function).

This notion, first introduced by Shannon is his 1959 paper, "Coding theorems for a discrete Soura with a fidelity criterion", would appear to reconcile the two problems of communicating information accurately and communicating it efficiently, given a Pay-off Function.

We will define this function formally. Its interpretation will appear immediately and justify why, intuitively, it had to be considered.

Let $E = \{1,\cdots,g,\cdots,G\}$ be the set of events (or messages) and $A = \{1,\cdots,k,\cdots,H\}$ be the set of actions. Let $(E,w(a|e),A)$ be an arbitrary channel with input alphabet E and output alphabet A. Let $d(e,a)$ be the loss function and $P(\cdot)$ the probability distribution on the messages generated by the source.

Consider

$$(1)\quad d(w(\cdot|\cdot)) = E_{e,a}\{d(e,a)\} = \sum_{e,a} P(e)\, w(a|e)\, d(e,a)$$

$$(2)\quad I(w(\cdot|\cdot)) = \sum_{e,a} P(e)\, w(a|e) \log \frac{w(a|e)}{\sum_{a'} P(e)w(a'|e)}$$

By definition:

$$R(D) = \inf_{w(\cdot|\cdot)} I(w(\cdot|\cdot))$$

with the constraint

$$d(w(\cdot|\cdot)) \leq D.$$

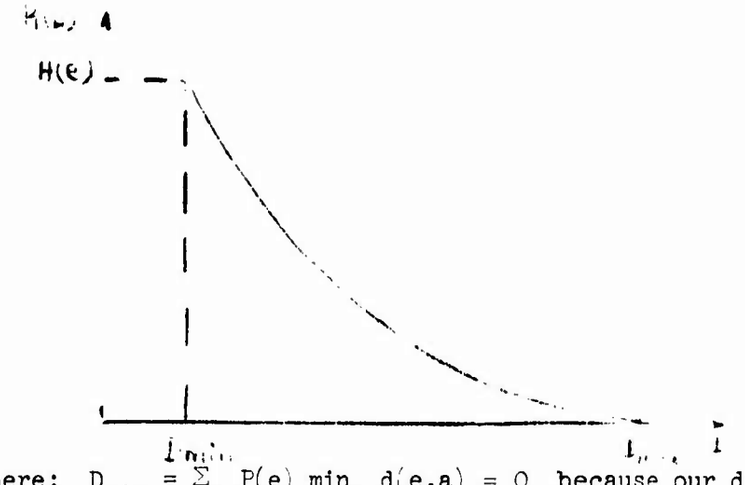
Note that $I(\cdot)$ is a continuous function of $w(\cdot|\cdot)$ and that the domain of $w(\cdot|\cdot)$ is closed and bounded. The inf is in fact a minimum when it exists. Moreover, $R(D)$ is decreasing in D since as D increases the domain of minimization increases. One shows quite easily that $R(D)$ is convex downward and that the constraint $d(w(\cdot|\cdot)) \leq D$ is equivalent to $d(w(\cdot|\cdot)) = D$.

$d(w(\cdot|\cdot))$ is a measure of the average loss, $I(w(\cdot|\cdot))$ is the average rate of information through $(E,w(a|e),A)$. This last quantity is proportional to the effort we must make to transmit the messages. We would like to make both of these quantities as small as possible, which of course is not feasible. So, given the source and $P(\cdot)$ and $d(\cdot,\cdot)$, it is important to know what is the smallest rate of information consistent with the maintenance of a loss no greater than some specified level, or equivalently, what is the smallest loss we can achieve if the rate is fixed.

The answers to these questions are given by $R(D)$, the so-called Rate-Distortion function, or Information rate of the Source for a loss level D. (This has to be proved because the minimum was taken over a very restricted class of information systems.) Shannon's coding theorem states that, with some mild restriction on $P(\cdot)$ and $d(\cdot,\cdot)$, $R(D)$ is the minimum achievable rate of information consistent with

$E_{e,a}\{d(e,a)\} \leq D$. Or, for any $\varepsilon > 0$, there exist codes with $E_{e,a}\{d(e,a)\} \leq D$ and Rate $\leq R(D) + \varepsilon$. Conversely, there exists no code with average loss D and rate less than $R(D)$.

Typically, $R(D)$ is found to have the following general shape:

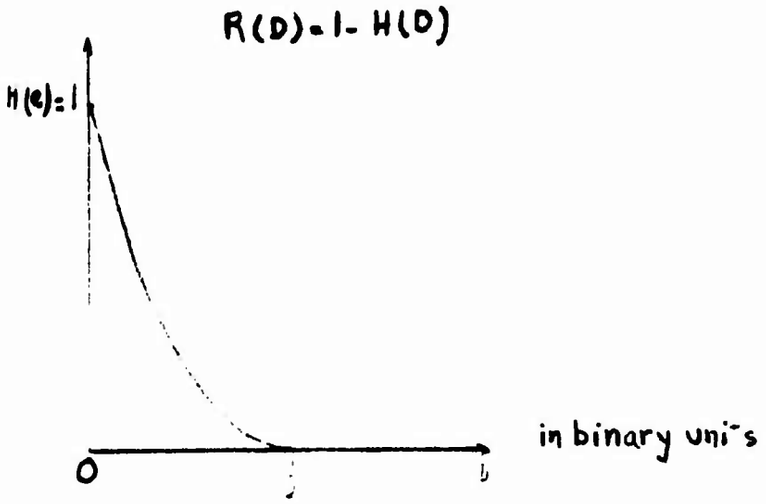


where: $D_{min} = \Sigma_e P(e) \min_a d(e,a) = 0$ because our distortion function has the property that for any e $\exists$ an a such that $d(e,a) = 0$. This point is achieved with a deterministic channel with $w(a(e)|e) = 1$ and $w(a|e) = 0$ for all $a \neq a(e)$. The corresponding value $R(0)$ of $R(D)$ is the entropy $H(e)$ of the source.

$D_{max} = \min_a \Sigma_e P(e)\, d(e,a)$ is the minimum achievable average loss with no information. Here the $w(a|e)$ matrix has a column of ones, all the other entries being zero. The capacity of such a channel is null because it has identical rows (see Ash, for example).

For a binary, uniform, memoryless source, with the loss function:

$$d(e,a) = 1 - \Delta(e,a) = \begin{cases} 0 & \text{if } e = a \\ 1 & \text{if } e \neq a. \end{cases}$$



2.2. Notation and Definitions

2.2.1. The source s produces a sequence $\{e_k\}_{k=1}^{\infty}$ of messages (or events) at a fixed rate of 1 message per second, each e_k being taken at random from a finite set $E = \{1,\cdots,g,\cdots,G\}$. The process $\{e_k\}_{k=1}^{\infty}$ is a sequence of independent, identically distributed random variables. Prob $\{e_k = g\} = P(e = g) = P(e)$ $\forall$ $k = 1,2,\cdots$.

2.2.2. The channel K is discrete, memoryless with input alphabet $X = \{1,\cdots,i,\cdots,I\}$, output alphabet $Y = \{1,\cdots,j,\cdots,J\}$, Prob $\{y = j|x = i\} = p(y|x)$. The channel capacity per use is C, and it can be used at most once every second.

2.2.3. The actions form a sequence $\{a_k\}_{k=1}^{\infty}$, each a_k is chosen within a finite set $A = \{1,\cdots,k,\cdots,H\}$.

2.2.4. Blocks of n events $e^n = e_1 \cdots e_k \cdots e_n \in E^n = \overbrace{E \times E \cdots \times E}^{n}$ are encoded into blocks of n channel input letters:

$$x^n = x_1 \cdots x_k \cdots x_n \in X^n.$$

Blocks of n output letters are recovered:

$$y^n = y_1 \cdots y_k \cdots y_n \in Y^n,$$

which are decoded into blocks of n actions:

$$x^n = x_1 \cdots x_k \cdots x_n \in A^n.$$

REMARK: As was said in the introduction, it is not efficient to restrict ourselves to vocabularies where all words have equal length.

2.2.5. A code of length H consists of: a vocabulary μ of M channel words of length n:

$$\mu = \{u_1, \cdots, u_m, \cdots, u_M\} \quad u_m \in X^n$$

and of two functions respectively called Encoding and Decoding functions:

$$\psi_1 : E^n \to \mathcal{U} \subset X^n$$

$$\psi_2 : Y^n \to \mathcal{V} \subset A^n$$

2.2.6. The rate of a code (ψ_1,ψ_2) is defined to be $\psi = \frac{1}{n} \log M$ where the log is of base $\underline{e}$. We will sometimes use $\log_2$ to express final results because a bit of information is more readily interpretable than a nat. It suffices though to remember that 1 nat = $\frac{1 \text{ bit}}{\log_2 \underline{e}}$.

2.2.7. The loss-measure. We recall that the loss when event e has occurred and action a has been taken is defined to be:

$$d(e,a) = \omega(e,a(e)) - \omega(e,a)$$

$$\omega(e,a(e)) \geq \omega(e,a) \quad \forall \ a,$$

where $\omega(e,a)$ is a finite Pay-off Function. $d(e,a)$ is a non-negative function of e and a and, for e fixed, it assumes the value zero at least once.

By definition:

$$d(e^n,a^n) \triangleq \frac{1}{n} \sum_{k=1}^{n} d(e_k,a_k)$$

This definition implies that time does not enter this problem. It intervenes only through the message and decision rates.

The treatment of time would introduce more parameters such as: encoding time, transmission time, decoding time, discount factor,....

The overall loss for the system of information $(\psi_1(\cdot),T(\cdot),\psi_2(\cdot))_n$ is then

$$d = \sum_{e^n,a^n} P(e^n)\ \mathrm{Prob}_{\psi_1,T,\psi_2}(a^n|e^n)\ d(e^n,a^n).$$

$\mathrm{Prob}_{\psi_1,T,\psi_2}(a^n|e^n)$ is the result of the composition of $\psi_1(\cdot),T(\cdot)$ and $\psi_2(\cdot)$ in that order.

2.2.8. Processing and Transmission Loss Functions. We have to cope with a major difficulty in connection with $\psi_1(\cdot)$: a channel input letter cannot be loaded, on the average, with an amount of information larger than or equal to the channel capacity, C, because the information which was loaded will eventually be entirely lost after transmission. If $H(e) \geq C$, one is forced to resort to what transmission engineers call a noisy code, ie., a code where the same word may represent several messages. In doing so, we present the channel with coarser information, that is, information of lesser value. The information is coarsened to the extent necessary so that it can be carried through the channel.

Practically, an efficient choice of $\psi_1(\cdot)$ and $\psi_2(\cdot)$ is possible only if one has a measure of the loss (of information value)

due to $\psi_1(\cdot)$, and a measure of the loss due to $T(\cdot)$ and $\psi_2(\cdot)$ such that the total loss is given by processing $(\psi_1(\cdot))$ + transmitting $(T(\cdot) * \psi_2(\cdot))$ losses.

Unless there is a well-defined relationship between channel input letters and actions, it seems difficult to derive these measures from $d(.,.)$. For the sake of simplicity, we have assumed $I = H$ (i.e., the channel input alphabet and the action set are of equal size), which lead us to consider all possible 1-1 correspondences between X and A. Not making this assumption would bring about more complex associations, but the increased difficulty, we think, is not insurmountable.

Let momentarily denote x_a the x associated with a particular a:

DEFINITION. $\delta_1(e,x_a) \stackrel{\Delta}{=} d(e,a)$ is called Processing Loss Measure (Function) $\delta_1(e^n,x^n) \stackrel{\Delta}{=} \frac{1}{n} \sum_{k=1}^{n} \delta_1(e_k,x_k)$.

The average Processing Loss, δ_1, equals $\sum_{e^n} P(e^n)\, \delta_1(e^n,\psi_1(e^n))$.

The transmission loss when x^n is sent and x'^n is received is given by

$$\sum_{e^n \in \psi_1^{-1}(x^n)} P(e^n)[\delta_1(e^n,x'^n) - \delta_1(e^n,x^n)].$$

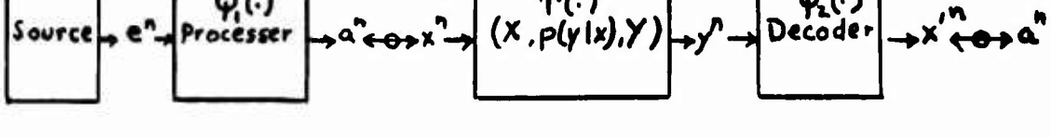
However, the proof of our transmission loss bound required a single letter measure (i.e., $\delta_2(x'^n,x^n) = \frac{1}{n} \sum_{k=1}^{n} \delta_2(x'_k,x_k)$). That brought about the following definition.

DEFINITION. $\delta_2(x,x') \stackrel{\Delta}{=} \max \{\Sigma_e P(e)\chi_{xx'}(e)\, \delta_1(e,x') - \delta_1(e,x)];$ $\Sigma_e P(e)\chi_{x'x}(e)\, [\delta_1(e,x) - \delta_1(e,x')]\}$, where $\chi_{xx'}(e)$ is the indicator function of $\{e : \delta_1(e,x') > \delta_1(e,x)\}$.

It is justified by the fact that

$$\frac{1}{n}\sum_{k=1}^{n} \delta_2(x_k,x'_k) \geq \sum_{e^n \epsilon \psi_1^{-1}(x^n)} P(e^n)[\delta_1(e^n,x'^n) - \delta_1(e^n,x^n)].$$

2.2.9. Transmission Scheme.



Processing Loss = $E_{e^n}\{\delta_1(e^n,\psi_1(e^n))\} = \delta_1$ $\leq$ Transmission Loss = $E_{x^n,y^n}\{\delta_2(x^n,\psi_2(y^n))\} = \delta_2$

$$\text{Total loss} \leq E_{e^n}\{\delta_1(e^n,\psi_1(e^n))\} + E_{x^n,y^n}\{\delta_2(x^n,\psi_2(y^n))\}$$

SECTION 3

The lower bound on Communication Loss for a given channel capacity

The theorem we are about to state is due to Shannon (1959). It answers the question, 'What is the smallest average loss one should expect with a channel of capacity C?'

THEOREM I. For an independent memoryless source S with a finite loss function $d(\cdot,\cdot)$, and a discrete memoryless channel of capacity C, there exists no encoding and decoding function that yields an average loss smaller than $R^{-1}(C)$, where $R^{-1}(\cdot)$ is the inverse of the Rate-loss function $R(\cdot)$.

In other words, for any code (ψ_1,ψ_2) the average loss, $D \geq R^{-1}(C)$. We will give a condensed version of Shannon's proof:

Suppose D is the average loss for (ψ_1,ψ_2), a block code of length n. $nC \geq I(x^n;y^n)$ by definition of capacity for a discrete memoryless channel $I(x^n;y^n) \geq I(e^n;a^n)$ by the data processing theorem (Feinstein S.3.3) $I(e^n;a^n) = H(e^n) - H(e^n|a^n) \geq \sum_{k=1}^{n} [H(e_k) - H(e_k|a_k)]$ because $H(e^n|a^n) = H(e_1|a_1 \cdots a_n) + H(e_2|e_1,a_1 \cdots a_n) + \cdots + H(e_n|e_1,\cdots,e_{n-1},a_1,\cdots,a_n) \leq H(e_1|a_1) + \cdots + H(e_n|a_n)$.

Now $\Sigma_{i=1}^{n} I(e_k,a_k) > nR(D)$ by definition of $R(D)$

$\therefore\ C > R(D)\ \ \therefore D > R^{-1}(C)$ because $R(\cdot)$ is decreasing.

q.e.d.

COROLLARY I. For an independent memoryless source s with a finite loss function $d(\cdot,\cdot)$, there exists no encoding function that yields a processing loss smaller than $R^{-1}(H(x))$ if $H(x)$ is the entropy of the channel input letters in the encoded messages.

Proof. This result clearly holds when the channel function $T(\cdot)$ and $\psi_2(\cdot)$ are identity transformations.

$$I(x^n;y^n) = H(x^n) \triangleq H(x)$$

$\therefore$ the same string of inequalities yields

$$C > H(x) > R(D)\ \ \therefore\ \ D > R^{-1}(H(x)) \qquad \text{q.e.d.}$$

In the next section, we will prove that there exist $\psi_1(\cdot)$'s such that

$$\delta_1 \leq D \qquad 0 \leq D \leq D_{max}$$

$$H(x) \leq R(D) + \varepsilon \qquad \forall\ \varepsilon > 0$$

SECTION 4

4.1. The Processing Loss Theorem. An Upper Bound to Processing Loss

We are faced here with the problem of processing the messages from the source (source encoding) in such a way as to decrease the rate of information to be sent through the channel so that there will be a least possible loss in information value for the user. The relevant loss measure is $\delta_1(\cdot,\cdot) : E \times X \to \text{Reals}$. We recall that a 1-1 correspondence was established between A and X and that $\delta_1(e,x)$ was defined to be equal to $d(e,a)$ for the associated couple (x,a). It follows that the rate loss function for the source δ, with $\delta_1(\cdot,\cdot)$, $R_1(A_1)$, is identical to $R(\cdot)$.

The encoding procedure we will describe is due to Shannon for its basic idea, which was later improved by Yudkin, and to Jelinek for its final version.

The problem was: given a memoryless source, governed by a distribution $P(\cdot)$ over the outputs $e \in E$, an alphabet $X = \{1,\cdots,H\}$ and a loss measure $\delta_1(\cdot,\cdot) : E \times X \to \text{Reals}$, let $\psi_1(\cdot)$ denote an encoding function that maps sequences e^n onto a set, $U = \{u_1,\cdots,u_M\}$ of M sequences of letters, $u_m = \{x_{1m},\cdots,u_{nm}\}$, called a vocabulary. What is the least obtainable value of the average loss $E_{e^n}\{\delta_1(e^n,\psi_1(e^n))\}$ and what is the corresponding $\psi_1(\cdot)$?

The only general answer to this question is obtained through a "random coding argument", which involves "threshold type encoders" proved to be efficient for large n's. In Shannon's approach, the threshold was a constant. In Yudkin's, it is a function of the sequence e^n designed in such a way as to minimize the expected processing loss. We may think of $\delta_1(e,x)$ as a distance between e and x (although it does not necessarily have the properties of a mathematical distance) and say that an optimal $\psi_1(\cdot)$ should map each sequence e^n onto its closest code word in the vocabulary.

We will now reproduce the main steps of Jelinek's proof. We will need these in Section 4.2.

$\psi_1(\cdot)$ is defined in the following fashion, given u:

$$\begin{cases} \psi_1(e^n)_u = u_m \text{ if } \delta_1(e^n,u_m) > d_0(e^n) \text{ for } m' = 1,\cdots,m-1 \\ \qquad\qquad \text{and } \delta_1(e^n,u_m) \leq d_0(e^n) \\ \psi_1(e^n)_u = u_M \text{ otherwise.} \end{cases}$$

$d_0(e^n)$ is a function over E^n whose exact form will be determined later according to the statistical properties of the source and the loss function.

Let d_{max} be the largest element in the loss matrix $[\delta_1(e,x)\delta]$ and $\phi_1(\cdot)$ the characteristic function defined on E^n as follows:

$$\phi_1(e^n)_u = \begin{cases} 1 & \text{if } \delta(e^n,u_m) > d_0(e^n) \text{ for all } m = 1,\cdots,M \\ 0 & \text{otherwise.} \end{cases}$$

The expected loss, given u, and $d_0(\cdot)$, can be bounded:

$$E_{e^n}\{\delta_1(e^n,\psi_1(e^n)_u)\} = \delta_1 \leq \sum_{e^n} P(e^n)\, d_0(e^n)[1 - \phi_1(e^n)_u] + d_{max} \sum_{e^n} P(e^n)\, \phi_1(e^n)_u \tag{4.1.1}$$

In order to estimate the bound, a random coding argument is resorted to. Let us assume that the code words u_m of u are selected independently at random, with probability

$$\text{Prob}\{u_m = x_m^t\} = \prod_{k=1}^{n} q(x_{k,m}) \triangleq Q(u_m)$$

$$\text{Prob}\{u\} = \prod_{m=1}^{n} Q(u_m),$$

where $q(\cdot)$ is an arbitrary distribution on X. Then:

$$\bar{\delta}_1 \triangleq E_u\{E_{e^n}[\delta_1(e^n,\psi_1(e^n)_u)]\} \leq \sum_{e^n} P(e^n) d_0(e^n) + d_{max} \sum_{e^n} P(e^n)\{P_q\{x^n : \delta_1(e^n,x^n) > d_0(e^n)\}\}^M \tag{4.1.2}$$

The M^{th} power appears because the M codewords are selected independently.

Now, by the inequality $\log x \leq x - 1$ or $x \leq e^{x-1}$, the above inequality can be written:

$$\bar{\delta}_1 \leq \sum_{e^n} P(e^n) d_0(e^n) + d_{max} \sum_{e^n} P(e^n) \exp\left[-P_q\{x^n : \delta_1(e^n,x^n) \leq d_0(e^n)\}\right] \cdot M \tag{4.1.3}$$

Let $J \triangleq \{e^n: P_q\{\delta_1(e^n,x^n) \leq d_0(e^n)\} < \frac{n}{M}\}$.

Then:

$$\sum_{e^n} P(e^n)(P_q(\delta_1(e^n,x^n) \leq d_0(e^n))^M \leq \sum_{e^n \in J} P(e^n) + \sum_{e^n \in J^C} P(e^n) e^{-n}$$

$$\leq e^{-n} + \sum_{e^n} P(e^n) \left[\frac{ne^{-n\psi}}{P_q\{\delta_1(e^n,x^n) \leq d_0(e^n)\}}\right]^{\sigma};$$

$$\sigma \geq 0 \tag{4.1.4}$$

The last term is obtained by using the usual bound to the indicator function of J and the definition of the rate of a code $\psi \triangleq \frac{1}{n} \log M$.

A lower bound to the probability in the denominator is found in Chapter 8 of Fano's, "Transmission of Information".

It can be shown that:

$$P_q\{\delta_1(e^n,x^n) \leq \frac{1}{n}\gamma'_n(\rho,e^n)\} \geq \exp\,[\gamma_n(\rho,e^n) - \rho\gamma'_n(\rho,e^n) + \frac{GH}{2}\log 2\pi n - B(\rho,e^n)] \qquad -\infty < \rho < 0 \tag{4.1.5}$$

Where: $-B(\rho,e^n)$ is a monotonically decreasing function of ρ, is independent of n and is bounded for all e^n's, as long as $|\rho|,G,H$ are finite.

$$-\gamma_n(\rho,e^n) = \sum_{k=1}^{n}\gamma(\rho,e_k) \triangleq \sum_{k=1}^{n}\log\,[\sum_x q(x)\,\underline{e}^{\rho\delta_1(e_k,x)}] \tag{4.1.6}$$

If we choose $d_0(e^n) = \frac{1}{n}\gamma'_n(\rho,e^n)$, (3.1.3) becomes

$$\delta_1 \leq \sum_{e^n} P(e^n)\,\frac{1}{n}\gamma'_n(\rho,e^n) + d_{max}\,[\underline{e}^{-n} + n^{\frac{GH}{2}}\beta(\rho)\underline{e}^{-n[\sigma\psi-\mu(\sigma)]}] \qquad -\infty < \rho < 0 \qquad 0 \leq \sigma \leq 1 \tag{4.1.7}$$

Where: $\mu(\sigma) \triangleq -\log\,[\sum_e P(e)\,e^{-\sigma(\gamma(\rho,e)-\rho\gamma'(\rho,e))}]$

$$-\beta(\rho) \triangleq (2\pi)^{\frac{H(G-1)}{2}}\exp\,[\max_{e^n} B(\rho,e^n)]$$

Finally, provided $\exists\ \sigma \in [0,1]$ such that $\mu'(\sigma) = \psi$

$$\bar{\delta}_1 \leq \sum_e P(e)\gamma'(\rho,e) + d_{max}\,[\underline{e}^{-n} + n^{\frac{GH}{2}}\beta(\rho)\underline{e}^{-n(\sigma\mu'(\sigma)-\mu(\sigma))}] \qquad -\infty < \rho < 0 \tag{4.1.8}$$

Let us define:

$$w_\rho(x|e) = \frac{q(x)\underline{e}^{\rho\delta_1(e,x)}}{\Sigma_{x'\in X}\, q(x')\underline{e}^{\rho\delta_1(e,x')}}$$

so that

$$\gamma'(\rho,e) = \sum_x w_\rho(x|e)\delta_1(e,x)$$

and

$$\overline{\delta}_1 \leq \sum_e P(e) \sum_x w_\rho(x|e)\delta_1(e,x) + d_{max}\left[\underline{e}^{-n} + n^{\frac{GH}{2}}\beta(\rho)\underline{e}^{-n(\sigma\mu'(\sigma)-\mu(\sigma))}\right]$$

The second term tends to zero with n as long as $\sigma\psi - \mu(\sigma) > 0$. We want to minimize ψ under the constraint:

$$\sum_{e,x} P(e)\, w_\rho(x|e)\delta_1(e,x) \leq \Delta_1 \qquad \Delta_1 \in [0, D_M]$$

$$\begin{cases} \psi_{min}(\Delta_1) = \min_{(\rho,q(\cdot))\in\Lambda(\Delta_1)} \left[\sum_{e,x} P(e)\, w_\rho(x|e) \log \frac{w_\rho(x|e)}{q(x)}\right] \\ \Lambda(\Delta_1) \triangleq \{(\rho,q(\cdot)) : \sum_{e,x} P(e)\, w_\rho(x|e)\delta_1(e,x) \leq \Delta_1;\ \rho \leq 0\} \end{cases} \tag{4.1.9}$$

It turns out that $\psi_{min}(\Delta_1) = R(\Delta_1)$ (4.1.10) for a complete proof, I refer to Jelinek, Chap. 11 section 11.3 and 11.4.

THEOREM 11. Let S, a constant, memoryless source governed by a distribution $P(\cdot)$ over the messages $e \in E$, a loss matrix $[\delta_1(e,x)]$, $e \in E$, $x \in X$, and a number $\Delta_1 \in [0,D_M]$ be given. The random family of encoding functions $\psi_1(\cdot)_u$, we have considered, of output sequences e^n onto a set $u = \{u_1,\ldots,u_M\}$ of $M = e^{n\psi}$ codewords $u_m = (x_{m1},\ldots,x_{mn})$ yield on the average a processing loss $\bar{\delta}_1$ such that:

$$\bar{\delta}_1 \triangleq E_{e^n u}\{\delta_1(e^n,\psi_1(e^n)_u)\} \leq \Delta_1 + d_{max}\{e^{-n} + n^{\frac{GH}{2}}\beta(\rho)e^{-n\alpha(\psi,\Delta_1)}\} \tag{4.1.11}$$

Where

d_{max} is the maximal entry in $[\delta_1(e,x)]$

$$\sum_{e,x} P(e)\, w_\rho(x|e)\, \delta_1(e,x) = \Delta_1$$

$$w_\rho(x|e) = \frac{q(x)e^{\rho\delta_1(e,x)}}{\Sigma_{x'} q(x')e^{\rho\delta_1(e,x')}}$$

and

$$\alpha(\psi,\Delta_1) > 0 \quad \text{provided} \quad \psi > R(\Delta_1).$$

COROLLARY II. There exists an encoding function of sequences e^n onto a vocabulary $u = \{u_1,\ldots,u_M\} \subset X^n$, $M = e^{n\psi}$ whose loss, δ_1, is bounded by the same expression.

4.2. Derivation of the Communication Loss Measure (Function) $\delta_2(x,x')$

Let us come back a few steps from the point we have reached. Let $q(\cdot)$ and $\rho < 0$ be arbitrary. Let ψ be equal to $\mu'(\sigma)$, $0 < \sigma < 1$.

Let $\mathcal{U}$ be a vocabulary of $e^{n\psi}$ codewords ($\psi > \Sigma_{e,x} P(e)w_\rho(x|e) \log \frac{w_\rho(x|e)}{q(x)}$) $\mathcal{U} = \{u_1, \cdots, u_M\}$. The encoding rule is the following:

Encode e^n as u_m if u_m is the first code word in $\mathcal{U}$ such that $\delta_1(e^n,u_m) \leq d_0(e^n)$; if this never occurs, encode e^n as u_M, or, which is the same, do not encode it at all. We understand how this decreases the quantity of information to be communicated it is now $\leq \psi$ nats, on the average, per event. From (3.1.1)

$$E_{e^n} \{\delta_1(e^n,\psi_1(e^n)_{\mathcal{U}})\} \leq \sum_{e^n} P(e^n)d_0(e^n) + d_{max} \sum_{e^n} P(e^n)\phi_1(e^n)_{\mathcal{U}} \tag{4.2.1}$$

The second term overbounds the processing loss due to not encoding certain e^n's. The first overbounds the loss due to the actual encoding of the other e^n's. Note that this term does not depend on the vocabulary $\mathcal{U}$, but only on $d_0(e^n)$ which, in turn, depends only on $q(x)$ and ρ.

Case of block length one:

There are $M = e^{\Psi}$ words in the vocabulary, chosen randomly with probability $q(x)$. A single message e is mapped on $u_m = x$ only if

$$\delta_1(e,x) \leq d_0(e) = \sum_x w_\rho(x|e)\, \delta_1(e,x).$$

What is the added loss if $u_m = x$ is decoded as x'?

If $\delta_1(e,x') \leq d_0(e)$ there might be a loss, but it has already been taken into account in the (source encoding bound) Processing loss bound. We need only be concerned about those e's such that $\delta_1(e,x) \leq d_0(e)$ and $\delta_1(e,x') > d_0(e)$. Clearly

$$\{e : \delta_1(e,x) \leq d_0(e) \text{ and } \delta_1(e,x') > d_0(e)\} \subset \{e : \delta_1(e,x) < \delta_1(e,x')\}.$$

Let $\chi_{xx'}(e)$ be the indicator function of $\{e : \delta_1(e,x) < \delta_1(e,x')\}$. Then the loss $\ell_2(x,x')$ due to Transmitting and decoding input x into x'

$$\ell_2(x,x') \leq \sum_e P(e)\chi_{xx'}(e)\, (\delta_1(e,x') - \delta_1(e,x)) \qquad (4.2.2)$$

DEFINITION. The transmission distortion function is

$$\delta_2(x,x') = \max \{ \sum_e P(e)\chi_{xx'}(e)(\delta_1(e,x') - \delta_1(e,x)); \quad (4.2.3)$$
$$\sum_e P(e)\chi_{x'x}(e)(\delta_1(e,x) - \delta_1(e,x'))\}.$$

This definition is unnatural a priori. The consideration of $\delta_2(.,.)$ is justified by:

- $\delta_2(\cdot,\cdot)$ is not function of the particular code at hand whereas $\ell_2(\cdot,\cdot)$ is. It is function of the given of this problem: $P(\cdot)$ and $d(\cdot,\cdot)$.

- $\ell_2(x,x') \leq \delta_2(x,x') \quad \forall\, x$ and x'.

- $\ell_2(x^n,x'^n) \leq \frac{1}{n} \sum_{k=1}^{n} \delta_2(x_k,x'_k)$. This last property is crucial in the proof of theorem III. It will be proved below.

Case of block length n:

LEMMA. The loss $\ell_2(x^n,x'^n)$ due to transmitting and decoding channel input x^n into x'^n is over-approximated by the single letter loss function $\delta_2(x,x')$:

$$\ell_2(x^n,x'^n) \leq \frac{1}{n} \sum_{k=1}^{n} \delta_2(x_k,x'_k).$$

Proof.

$$\ell_2(x^n,x'^n) \leq \sum_{e^n \in \mathcal{E}^n} P(e^n)[\delta_1(e^n,x'^n) - \delta_1(e^n,x^n)]$$

where:

$$\mathcal{E}^n = \{e^n : \delta_1(e^n,x^n) < \delta_1(e^n,x'^n)\}$$

now:

$$\sum_{e^n \in \mathcal{E}^n} P(e^n)[\delta_1(e^n,x'^n) - \delta_1(e^n,x^n)]$$

$$= \sum_{e^n \in \mathcal{E}^n} P(e^n) \frac{1}{n} \sum_{k=1}^{n} (\delta_1(e_k,x'_k) - \delta_1(e_k,x_k))$$

I claim that

$$\sum_{k=1}^{n} (\delta_1(e_k,x'_k) - \delta_1(e_k,x_k))$$

$$\leq \sum_{k=1}^{n} (\delta_1(e_k,x'_k) - \delta_1(e_k,x_k))\chi_{x_k x'_k}(e_k)$$

Indeed, suppose that for a particular e_k, $\chi_{x_k x'_k}(e_k) = 0$. Then $\delta_1(e_k,x'_k) \leq \delta_1(e_k,x_k)$ and therefore, to a non-positive term in the first member corresponds a zero term in the second.

$$\therefore \ell_2(x^n,x'^n) \leq \sum_{e^n} P(e^n) \frac{1}{n} \sum_{k=1}^{n} \chi_{x_k x'_k}(e_k)(\delta_1(e_k,x'_k) - \delta_1(e_k,x_k))$$

Now, since $P(e^n) = P(e_1) \cdots P(e_n)$

$$\ell_2(x^n,x'^n) \leq \frac{1}{n} \sum_{k=1}^{n} \sum_{e} P(e) \chi_{x_k x'_k}(e)(\delta_1(e,x'_k) - \delta_1(e,x_k))$$

$$\therefore \ell_2(x^n,x'^n) \leq \frac{1}{n} \sum_{k=1}^{n} \delta_2(x_k,x'_k) \quad \text{by (3.2.3)} \qquad \text{q.e.d.}$$

REMARK: $\delta_2(x,x')$ is in fact the product of the probability of x being sent given that x belongs to the vocabulary u and the loss when the output of the channel is decoded into x'. Therefore, to compute the Expected transmission loss, we need only sum up the expected losses for each word in the vocabulary.

Properties of $\delta_2(\cdot,\cdot)$:

1) $\delta_2(x,x') \geq 0$, $\delta_2(x,x') = 0$ when $x = x'$

2) the $[\delta_2(x,x')]$ matrix is square and symmetric, an important property for what follows.

3) $\delta_2(\cdot,\cdot)$ has the triangular inequality property.

LEMMA. $\delta_2(x_1,x_3) \leq \delta_2(x_1,x_2) + \delta_2(x_2,x_3)$.

Proof. Let us first show that

$$\sum_e P(e)\chi_{x_1x_3}(e)(\delta_1(e,x_3) - \delta_1(e,x_1))$$

$$\leq \sum_e P(e)\chi_{x_1x_2}(e)(\delta_1(e,x_2) - \delta_1(e,x_1)) \qquad (4.2.4)$$

$$+ \sum_e P(e)\chi_{x_2x_3}(e)(\delta_1(e,x_3) - \delta_1(e,x_2))$$

the first member can be written

$$\sum_e P(e)\chi_{x_1x_3}(e)[\delta_1(e,x_3) - \delta_1(e,x_2) + \delta_1(e,x_2) - \delta_1(e,x_1)]$$

Suppose $\chi_{x_1x_3}(e) = 1$, i.e., $\delta_1(e,x_3) \geq \delta_1(e,x_1)$. These cases are possible.

1) $\delta_1(e,x_3) \geq \delta_1(e,x_2) > \delta_1(e,x_1)$ the two members of the inequality (4.2.4) are equal,

2) $\delta_1(e,x_2) > \delta_1(e,x_3) > \delta_1(e,x_1)$ the right-hand side of (4.2.4) is larger because it has a zero term rather than the negative term, $\delta_1(e,x_3) - \delta_1(e,x_2)$.

3) $\delta_1(e,x_3) > \delta_1(e,x_1) \geq \delta_1(e,x_2)$ the right-hand side of (4.2.4) is larger for the same reason as in 2).

If $\chi_{x_1x_3}(e) = 0$ then the right-hand side of (4.2.4) is larger than or equal to the left-hand side, because it is non-negative. We would prove in the same fashion that:

$$\sum_{e} P(e)\chi_{x_3x_1}(e)(\delta_1(e,x_1) - \delta_1(e,x_3))$$

$$\leq \sum_{e} P(e)\chi_{x_3x_2}(e)(\delta_1(e,x_2) - \delta_1(e,x_3)) \qquad (4.2.4')$$

$$+ \sum_{e} P(e)\chi_{x_2x_1}(e)(\delta_1(e,x_1) - \delta_1(e,x_2))$$

therefore, from (4.2.3), (4.2.4) and 4.2.4'), it follows that:

$$\delta_2(x_1,x_3) \leq \delta_2(x_1,x_2) + \delta_2(x_2,x_3) \qquad \text{q.e.d.}$$

SECTION 5

An Upperbound to the Expected Transmission Loss, Given a Channel

Let us be given a channel $(X,p(y|x),Y)$, where:

$$X = \{1,\cdots,i,\cdots,H\}$$

$$Y = \{1,\cdots,j,\cdots,J\}$$

and a transmission loss matrix $\{\delta_2(x,x')\}$. We recall that $\delta_2(\cdot,\cdot)$ was induced by $\delta_1(\cdot,\cdot)$ through an arbitrary 1-1 correspondence between the action set A and X.

We want now to associate channel input and output letters in order to define a distortion measure between channel input and channel output letters. Let us call $a(y)$ the action associated with y determined in the following fashion:

$$\sum_x q(x)p(y|x)\delta_2(x,a(y)) \leq \sum_x q(x)p(y|x)\delta_2(x,x') \quad \forall\ x',$$

where $q(x)$ is the probability distribution used on the x's.

Define: $\delta_3(x,y) \triangleq \delta_2(x,a(y))$.

We will suppose, furthermore, that the correspondence between A and X has been done so as to minimize

$$\sum_{x,y} q(x)p(y|x)\delta_2(x,a(y)).$$

The channel is now adapted to the source and the loss function $d(\cdot,\cdot)$.

We now propose to ask the following question: Let $\mathcal{U}$ be a vocabulary of M sequences $(u_1,\cdots,u_M)$ of channel input letters of length n : $u_m = (x_{m1},\cdots,x_{mn})$ and let the source messages e^n be mapped on the codewords u_m's by the rule given in 4.1. The loss when u_m is transmitted and decoded into $u_{m'}$ is over-approximated by $\delta_2(u_m,u_{m'})$. What is the best decoding function $\psi_2(\cdot) : Y^n \to A^n$ and what is the least obtainable value of the expected transmission loss?

As a matter of fact, the only optimal decoding function is the one which maps an output sequence y^n onto the code word u_m that minimizes the quantity,

$$\sum_{m'=1}^{M} \Pr(u_{m'}|y^n)\delta_2(u_m,u_{m'}).$$

Unfortunately this decoding rule is hardly feasible in practice, especially for large n's, and moreover one does not know how to evaluate its performances in terms of a bound.

We will instead use a so-called "minimum distance" decoding function, defined as follows:

$$\psi_2(y^n)_{\mathcal{U}} = u_m \quad \text{if} \quad \delta_3(u_m,y^n) \leq \delta_3(u_{m'},y^n) \;\forall\; m' = 1,\cdots,M \qquad (5.1)$$

When y^n is received, $\psi_2(\cdot)_{\mathcal{U}}$ decode it as the code word u_m which is the less distorted with respect to y^n.

Given a vocabulary $\mathcal{U} = \{u_1,\cdots,u_M\}$, we would like to evaluate the expected value of the transmission loss function:

$$\sum_{m=1}^{M} \sum_{y^n} \delta_2(u_m,\psi_2(y^n)_{\mathcal{U}}) \Pr\{y^n|u_m\} \qquad (5.2)$$

As in Section 4.1, we will only be able to evaluate the expected value of this quantity over all $\mathcal{U}$'s $= \{u_1,\cdots,u_M\}$ generated at random as in 4.1. We will prove that if the code rate ψ is small enough, the expected transmission loss, averaged over all $\mathcal{U}$'s, is bounded from above by a function which tends to zero with $n \to \infty$.

Proof. Let $\mathcal{U}$ be generated at random, each word u_m being chosen independently with probability Prob $\{u_m = x^n\} = Q(x^n) = \prod_{x=1}^{n} q(x_k)$ where $x^n = (x_1,\cdots,x_k,\cdots,x_n)$ and $q(\cdot)$ is the same distribution as in 4.1.

Let $x^n = (x_1 \cdots x_n)$ be a code word and y^n be received when x^n has been sent. The probability that the distortion $\delta_3(x^n,y^n)$ be larger than some value ϕ is bounded from above using a result of Fano's, "Transmission of Information", Chapter 8.

$$\text{Prob}\ \{\delta_3(x^n,y^n) > g'(x^n,s)\} \leq \underline{e}^{-n[sg'(x^n,s)-g(x^n,s)]} \quad s \geq 0 \quad (5.3)$$

where:

$$\begin{cases} g'(x^n,s) \text{ is an increasing function of } s \text{ and} \\ g'(x^n,0) = \sum_{Y^n} P(y^n|x^n)\delta_3(x^n,y^n) \\ sg'(x^n,s) - g(x^n,s) = \sum_{k=1}^{n} \frac{1}{n} \sum_{Y} P_s(y|x_k) \log \frac{P_s(y|x_k)}{P(y|x_k)} \geq 0 \\ \qquad \text{increasing with } s \\ P_s(y|x_k) = \dfrac{\underline{e}^{s\delta_3(x_k,x)} P(y|x_k)}{\Sigma_{y'} P(y'|x_k)\underline{e}^{s\delta_3(x_k,y')}} \end{cases}$$

If $\delta_3(x^n,y^n) > g'(x^n,s)$ then the transmission loss is less than or equal to $\delta_{2\ \max}$, the maximal entry in the $[\delta_2(x,x')]$ matrix. If $\delta_3(x^n,y^n) \leq g'(x^n,s)$ then a transmitting and decoding loss smaller than or equal to $2g'(x^n,s)$ could occur only if there was another code word x'^n such that $\delta_3(x'^n,y^n) \leq g'(x^n,s)$, otherwise, no loss. The probability of such a situation depends on y^n and $g'(x^n,s)$ only, since the code words are chosen independently. This probability, averaged over all y^n's such that $\delta_3(x^n,y^n) \leq g'(x^n,s)$ is taken to be equal to the probability of another code word at a distance less than $g'(x^n,s)$ from x^n.

Let us now estimate this probability using once again one of

Fano's bound:

$$\text{Prob}\ \{x'^n : \delta_2(x^n,x'^n) \le g'(x'^n,s)\} \le \underline{e}^{-n\{\lambda(x_n)f'(x^n,\lambda(x_n))\}-f(x^n,\lambda(x^n))\}};\ \lambda(x^n) \le 0 \tag{5.4}$$

where:

$$\begin{cases} g'(x^n,s) = f'(x^n,\lambda(x^n)) \ \text{ or } \ \lambda(x^n) = 0 \ \text{ if } \ g'(x^n,s) > f'(x^n,0). \\ \\ f'(x^n,\lambda) \ \text{ is an increasing function of } \ \lambda \\ \\ f'(x^n,\lambda) \le f'(x^n,0) = \sum_{x^n} Q(x^n)\delta_2(x^n,x'^n);\ \lambda \le 0 \\ \\ \lambda f'(x^n,\lambda) - f(x^n,\lambda) = \sum_{k=1}^{n} \frac{1}{n} \sum_{x' \in X} Q_\lambda(x'|x_k) \log \frac{Q_\lambda(x'|x_k)}{q(x')} \ge 0 \ \text{ for } \ \lambda \le 0 \\ \\ \qquad \text{is a decreasing function of } \ \lambda \\ \\ Q_\lambda(x'|x_k) = \dfrac{\underline{e}^{\lambda\delta_2(x_k,x')}\, q(x')}{\Sigma_{x'}\, q(x')\underline{e}^{\lambda\delta_2(x_k,x')}} \end{cases}$$

The probability that $\exists$ no other code word in the sphere of radius $g'(x^n,s)$ about x^n:

$$1 - \pi \ge [1 - \underline{e}^{-n(\lambda(x^n)f'(x^n,\lambda(x^n))-f(x^n,\lambda(x^n))}]^{M-1} \ge [\cdots]^M$$

$\therefore$ the probability that $\exists$ at least anocher code word in this sphere:

$$\pi \leq 1 - [\cdots]^M$$

Now $(1 - x)^{\frac{1}{x}} \geq e^{\frac{1}{x}(-x - \frac{x^2}{2})} \geq e^{-\frac{3}{2}}$ for $0 \leq x \leq 1$

$$\therefore \pi \leq 1 - \exp\left[-\tfrac{3}{2} e^{-n(\lambda(x_n)f'(x^n,\lambda(x^n))-f(x^n,\lambda(x^n)))M}\right]$$

but $1 - e^{-x} \leq x \;\forall\; x$ and $M = e^{n\psi}$

$$\therefore \pi \leq \tfrac{3}{2} e^{-n[\lambda(x_n)f'(x_n,\lambda(x_n))-f(x_n,\lambda(x_n))-\psi]}$$

So the loss, averaged over all μ's that have x_n as one of the code words, is bounded by

$$\begin{aligned} &E_{\mu(x^n)y^n}\{\delta_2(x^n,\psi_2(y^n)_{\mu(x^n)}\} \\ &\quad \leq 3g'(x^n,s)e^{-n[\lambda(x^n)f'(x_n,\lambda(x_n))-f(x_n,\lambda(x_n))-\psi]} \qquad (5.5) \\ &\quad + \delta_{2\,\max} e^{-n(sg'(x^n,s)-g(x^n,s))} \qquad \lambda(x^n) \geq 0, \quad s < 0 \end{aligned}$$

It is left now to average the loss over all vocabularies and all code words. The probability of x^n being a code word is equal to

$$1 - [1 - Q(x^n)]^M \simeq Q(x^n).$$

$\therefore$ the average loss for μ, averaged over all u's

$$\underset{u,y^n,x^n}{E} (\delta_2(x^n,\psi_2(y^n)_{u}))$$

$$\leq \underset{x^n}{E} \{3g'(x^n,s)\} \underset{x^n}{E} \{\underline{e}^{-n[\lambda(x^n)f'(x^n,\lambda(x^n))-f(x^n,\lambda(x^n))-\psi]}$$

$$+ \delta_2 \max \underset{x^n}{E} \{\underline{e}^{-n(g'(x^n,s)-g(x^n,s))}\} \quad \lambda(x^n) \leq 0, \; s > 0 \tag{5.6}$$

Now:

$$\underset{x^n}{E} (3g'(x^n,s)) = 3 \sum_{x^n} Q(x^n) \frac{1}{n} \sum_{k=1}^{n} \sum_{y^n} P_s(y|x_k)\delta_3(x_k,y)$$

$$= 3 \sum_{x,y} q(x) \; P_s(y|x)\delta_3(x,y)$$

because $Q(x^n) = \prod_{k=1}^{n} q(x_k)$.

$$\underset{x^n}{E} \{\underline{e}^{-n[\lambda(x^n)\cdot f'(x^n,\lambda(x^n))-f(x^n,\lambda(x^n))}\}$$

$$= \sum_{\{x^n:\lambda(x^n)<0\}} Q(x^n)\underline{e}^{-n[\lambda(x^n)\cdot f'(x^n,\lambda(x^n))-f(x^n,\lambda(x^n))]} \tag{5.7}$$

$$+ \sum_{\{x^n:\lambda(x^n)=0\}} Q(x^n)\underline{e}^{0} \qquad \text{because} \quad \lambda f'(x^n,\lambda) - f(x^n,\lambda) = 0 \text{ for } \lambda = 0$$

$$\sum_{\{x^n:\lambda(x^n)=0\}} Q(x^n) = P_q\{x^n : \lambda(x^n) = 0\}$$

$$= P_q\{\sum_{x'^n} Q(x'^n)\delta_2(x^n,x'^n) \leq \sum_{y^n} P_s(y^n|x^n)\delta_3(x^n,y^n)\}$$

by (5.4). $Q(x'^n) = \prod_{k=1}^{n} q(x'_k)$. $\therefore$ as $n \to \infty$:

$$\begin{cases} \sum_{x'^n} Q(x^n)\delta_2(x^n,x'^n) \to \sum_{x,x'} q(x)q(x')\delta_2(x,x') \\ \sum_{y^n} P_s(y^n|x^n)\delta_3(x^n,y^n) \to \sum_{x,y} q(x)P_s(y|x)\delta_3(x,y) \end{cases}$$

But $\Sigma_{x,y}\, q(x)\, P_s(y|x)\delta_3(x,y) = \Sigma_{x,y}\, q(x)\, P_s(y|x)\delta_2(x,a(y))$ because of the association between Y and A, by this association:

$$\sum_{x,y} q(x)p(y|x)\delta_2(x,a(y)) < \sum_{xx'} q(x)q(x')\delta_2(x,x')$$

$\therefore$ since $\Sigma_{x,y}\, q(x)\, P_s(y|x)\delta_2(x,a(y))$ is an increasing function of s, we can choose s small enough to ensure

$$\sum_{x,y} q(x)\, P_s(y|x)\delta_2(x,a(y)) < \sum_{x,x'} q(x)q(x')\delta_2(x,x').$$

$\therefore$ for s small enough $\text{Prob}_q\{\Sigma_{x'^n} Q(x'^n)\delta_2(x^n,x'^n) \leq \Sigma_{y^n} P_s(y^n|x^n)\delta_3(x^n,y^n)\}$ tends to zero exponentially as $n \to \infty$.

Now, because of the continuity in λ of $\lambda f'(x^n,\lambda) - f(x^n,\lambda)$ and because of the continuity of the expectation operation, there exists a <u>maximal</u> $\lambda < 0$ such that:

$$E_{x^n}\{e^{-n[\lambda(x^n)f'(x^n,\lambda(x^n))-f(x^n,\lambda(x^n))]}\} \leq e^{-nE_{x^n}\{\lambda_s f'(x^n,\lambda_s)-f(x^n,\lambda_s)\}}$$

$$E_{x^n}\{f'(x^n,\lambda_s)\} = \sum_{x,x'} q(x')Q_{\lambda_s}(x|x')\delta_2(x,x') \geq \sum_{x,y} q(x)P_s(y|x)\delta_3(x,y)$$

$$E_{x^n}\{\lambda f'(x^n,\lambda) - f(x^n,\lambda)\} = E_{x^n}\{\frac{1}{n}\sum_{k=1}^{n}\sum_{x'} Q_\lambda(x'|x_k)\log\frac{Q_\lambda(x'|x_k)}{q(x')}\}$$

$$E_{x^n}\{\lambda f'(x^n,\lambda) - f(x^n,\lambda)\} = \sum_{x,x'} q(x)Q_\lambda(x'|x)\log\frac{Q_\lambda(x'|x)}{q(x')} \qquad (5.8)$$

Finally, in order that $E_{x^n}\{e^{-n[\lambda(x^n)f(x^n,\lambda(x^n))-f(x^n,\lambda(x^n))]+n\psi}\}$ tends to zero with $n \to \infty$, it suffices that

$$\psi < \sum_{xx'} q(x)Q_{\lambda_s}(x'|x)\log\frac{Q_{\lambda_s}(x'|x)}{q(x')} \qquad (5.9)$$

In the same fashion

$$E_{x^n}\{g'(x^n,s)\} = E_x\{g'(x,s)\}$$

and

$$E_{x^n}\{\underline{e}^{-n[sg'(x^n,s)-g(x^n,s)]}\} = E_x\{\underline{e}^{-n[sg'(x,s)-g(x,s)]}\}$$

We have proved, so far, that the expected transmission loss, averaged over all $\mu = \{u_1,\cdots,u_M\}$, $(M = \underline{e}^{n\psi})$

$$\left\{E_{\mu,y^n,x^n}\{\delta_2(x^n,\psi_2(y^n))\} \le 3E_x\{g'(x,s)\}\underline{e}^{-n[\Sigma_{x,x'}q(x)Q_{\lambda_s}(x'|x)\log\frac{Q_{\lambda_s}(x'|x}{q(x')}-\psi]}\right.$$

$$\left. + \delta_2 \max E_x\{\underline{e}^{-n[sg'(x,s)-g(x,s)]}\} \quad \text{with} \quad s > 0\right.$$

$$\left\{\sum_{xx'} q(x)Q_{\lambda_s}(x'|x)\delta_2(x,x') \ge \sum_{x,y} q(x)P_s(y|x)\delta_3(x,y)\right. \tag{5.10}$$

Let us ask now a question of theoretical importance (its economic relevance cannot be asserted unless the utility function of the information system and the various cost functions are given):

What is the supremum, ψ^*, of the permissible code rates?

A priori $\psi^* \le C$ the capacity of the channel.

Let us prove that $\psi^* \ge C$:

Proof.

$$\left\{\begin{array}{l} \psi^* = \max_{q(\cdot),\lambda} \sum_{x,x'} q(x)Q_\lambda(x'|x) \log \dfrac{Q_\lambda(x'|x)}{q(x')} \\ \\ \text{with} \\ \\ \sum_{xx'} q(x)Q_\lambda(x'|x)\delta_2(x,x') = \sum_{x,y} q(x)p(y|x)\delta_3(x,y) \end{array}\right. \tag{5.11}$$

by 3.9, 3.10 and the fact that $\Sigma_{x,y} q(x)P_s(y|x)\delta_3(x,y)$ is minimum for $s = 0$. $P_0(y|x) = p(y|x)$.

$$C = \max_{q(\cdot)} \sum_{x,y} q(x)p(y|x) \log \frac{p(y|x)}{\Sigma_x q(x)p(y|x)}$$

by convexity U of $-\log(\cdot)$

$$C \leq \max_{q(\cdot)} \sum_{x,y} q(x)p(y|x) \log \frac{p(y|x)}{q(x)} \tag{5.12}$$

(3.12) can be written:

$$C \leq \max_{q(\cdot)} \sum_{xx'} q(x) \sum_{\{y:a(y)=x'\}} p(y|x) \log \frac{p(y|x)}{q(x)}, \tag{5.13}$$

where $a(y)$ is the action or equivalently the channel input letter associated with y.

Let us denote by $Q^*(x'|x)$, the sum $\Sigma_{\{y:a(y)=x'\}} p(y|x)$. By

convexity $\cap$ of $\log(\cdot)$

$$\sum_{\{y:a(y)=x'\}} p(y|x) \log \frac{p(y|x)}{q(x)} \leq \sum_{\{y:a(y)=x'\}} p(y|x) \log \frac{Q^*(x'|x)}{q(x)} \quad (5.14)$$

From (3.13) and (3.14)

$$C \leq \max_{q(\cdot)} \sum_{xx'} q(x)Q^*(x'|x) \log \frac{Q^*(x'|x)}{q(x)} \leq \psi^*,$$

because:

$$\sum_{x'x} q(x)Q^*(x'|x)\delta_2(x,x') = \sum_{x,y} q(x)p(y|x)\delta_3(x,y)$$

by definition of $a(y)$ and $Q^*(x'|x)$. q.e.d.

$\therefore \psi^* = C.$

Let us summarize the results obtained:

THEOREM III. Let $(X,p(y|x),Y)$, a discrete memoryless channel of capacity C and a transmission loss function $\delta_2(\cdot,\cdot) : X \times X \rightarrow$ Reals be given. Let $q(\cdot)$ be an arbitrary probability distribution on X. Let $u = \{u_1,\cdots,u_m,\cdots,u_M\}$, $M = e^{n\psi}$, $u_m = (x_{m1},\cdots,u_{mn})$ be a vocabulary of code words of length n generated at random, each word being chosen independently with probability $Q(u_m) = \prod_{k=1}^{n} q(x_{mk})$. Then the expectation over all u's of the average transmission loss for a given vocabulary, when a "minimum distance" decoding function $\psi_2(\cdot) : Y^n \rightarrow A^n$ is used, satisfies the following inequality:

$$\bar{\delta}_2 = \underset{u,y^n,x^n}{E} \left(\delta_2(x^n,\psi_2(y^n)_u)\right) \leq 5\, E_x\,(g'(x,s))\underline{e}^{-n[\psi(\lambda_s)-\psi]}$$

$$+ \delta_{2\,\max}\, E_x\,(\underline{e}^{-n[sg'(x,s)-g(x,s)]}) \qquad s > 0 \tag{5.15}$$

Where $\delta_{2\,\max}$ is the $\max_{x,x'} \delta_2(x,x')$

$$\psi(\lambda_s) \triangleq \sum_{xx'} q(x)\, Q_{\lambda_s}(x'|x) \log \frac{Q_{\lambda_s}(x'|x)}{q(x)} > 0 \quad \text{for} \quad s > 0$$

$$sg'(x^n,s) - g(x^n,s) > 0 \quad \text{for} \quad s > 0$$

$$Q_{\lambda_s}(x'|x) = \frac{q(x')\underline{e}^{\lambda_s \delta_2(x,x')}}{\Sigma_{x'}\, q(x')\underline{e}^{\lambda_s \delta_2(x,x')}}$$

$$P_s(y|x) = \frac{p(y|x)\underline{e}^{s\delta_2(x,a(y))}}{\Sigma_{y'}\, P(y'|x)\underline{e}^{s\delta_2(x,a(y))}}$$

$a(y)$ is the action associated with output y

$$\sum_{xx'} q(x) Q_{\lambda_s}(x'|x)\delta_2(x,x') \geq \sum_{x,y} q(x)\, P_s(y|x)\delta_2(x,a(y))$$

$$\sup_{q(\cdot),s} \psi(\lambda_s) = C$$

and the bound tends to zero when $s > 0$ and $\psi < \psi(\lambda_s)$.

COROLLARY III. There exists a vocabulary $\mathcal{U} = (u_1, \cdots, u_M)$, $u_m = (u_{m1}, \cdots, u_{mn})$, $M = \underline{e}^{n\psi}$ whose average loss, δ_2, satisfies inequality (3.15).

SECTION 6

The Communication Loss Theorems

We recall that, for the sake of analysing the effects of encoding on the one hand and transmitting and decoding on the other hand, we introduced, respectively, the Processing Loss Function $\delta_1(\cdot,\cdot)$ and the Transmission Loss Function $\delta_2(\cdot,\cdot)$ in such a way that the expectation of the loss due to communication, d, be less than or equal to the sum of the expectation of the loss due to encoding, δ_1, and the expectation of the loss due to transmitting and decoding, δ_2.

$$d \leq \delta_1 + \delta_2 \tag{6.1}$$

In section four, we proved that:

- If a vocabulary $\mathcal{U} = \{u_1,\cdots,u_M\}$, $M = \underline{e}^{n\psi}$, $u_m = (u_{m1},\cdots,u_{mn})$ is generated at random with Prob $\{\mathcal{U}\} = \Pi_{m=1}^{M} \Pi_{k=1}^{n} q(x_{mk})$, $q(\cdot)$ being an arbitrary distribution on the channel alphabet X.

- If the blocks of source messages of length n are mapped onto $\mathcal{U}$ in such a way that $\psi_1(e^n)_{\mathcal{U}} = u_m$ only if $\delta_1(e^n,u_m) \leq d_0(e^n)$. Then the average processing loss $\bar{\delta}_1 = E_{\mathcal{U},x^n,y^n}\{\delta_1(e^n,\psi_1(e^n)_{\mathcal{U}})\}$ satisfies:

$$\bar{\delta}_1 \le \sum_e P(e)\gamma'(\rho,e) + d_{max}\left[\underline{e}^{-n} + n^{\frac{GH}{2}}\beta(\rho)\underline{e}^{-n(\sigma\psi-\mu(\sigma))}\right]$$

$$\rho < \infty \quad 0 < \sigma < 1$$

$$d_0(e^n) = \frac{1}{n}\sum_{k=1}^{n_1}\gamma'(\rho,e_k) \triangleq \frac{1}{n}\sum_{k=1}^{n_1}\sum_x W_\rho(x|e_k)\delta_1(e_k,x)$$

and $\sigma\psi - \mu(\sigma) > 0$ if $\psi = \mu'(\sigma) \ge \mu'(0) = \sum_{e,x} P(e)W_\rho(x|e)\log\frac{W_\rho(x|e)}{q(x)}$

In Section five, we proved that:

- If a vocabulary $\mathcal{U}$ is generated at random as in Section four, using a distribution $q(\cdot)$ on X, $\mathcal{U} = \{u_1,\cdots,u_M\}$, $M = e^{n\psi}$, $u_m = (x_{m1},\cdots,x_{mn})$.

- If the channel outputs of length n are decoded, using the "minimum distance" decoding function $\psi_2(\cdot)_{\mathcal{U}}$.

Then the expected transmission loss $\bar{\delta}_2 = E_{\mathcal{U},y^n,x^n}\{\delta_2(x^n,\psi_2(y^n)_{\mathcal{U}})\}$ satisfies:

$$\bar{\delta}_2 \le 3\{\sum_x q(x)g'(x,s)\}\underline{e}^{-n[\psi(\lambda_s)-\psi]} + \delta_2 \max\sum_x q(x)\underline{e}^{-n[sg'(x,s)-g(x,s)]}$$

where $\psi(\lambda_s) = \sum_{xx'} q(x)Q_{\lambda_s}(x'|x)\log\frac{Q_{\lambda_s}(x'|x)}{q(x)}$ (6.3)

and $\psi(\lambda_s) - \psi > 0$ if $\psi < 0$ if $\psi < \psi(\lambda_s)$

From (6.1), (6.2) and (6.3) it follows that:

THEOREM.

$$\bar{d} \leq \underbrace{\sum_e P(e)\gamma'(\rho,e)}_{(1)} \qquad \underbrace{\max}_{(2)} \left[\underline{e}^{-n} + n^{\frac{GH}{2}} \beta(\rho) \underline{e}^{-n(\sigma\psi - \mu(\sigma))} \right]$$

$$+ 3 \left\{ \sum_x q(x) g'(x,s) \right\} \underline{e}^{-n[\psi(\lambda_s) - \psi]} + \delta_{2\max} \sum_x q(x) \underline{e}^{-n[sg'(x,s) - g(x,s)]} \qquad (6.4)$$

$$\rho < 0, \quad 0 < \sigma < 1, \quad s > 0$$

where (2), (3) and (4) tend to zero as $n \to \infty$ if

$$\sum_{e,x} P(e) W_\rho(x|e) \log \frac{W_\rho(x|e)}{q(x)} \leq \psi < \sum_{xx'} q(x) Q_{\lambda_s}(x'|x) \log \frac{Q_{\lambda_s}(x'|x)}{q(x)} \qquad (6.5)$$

and $q(\cdot)$ is arbitrary.

Note that (6.5) can always be satisfied for $q(\cdot)$ and s given by taking $\rho < 0$ large enough because $\Sigma_{e,x} P(e) W_\rho(x|e) \log \frac{W_\rho(x|e)}{q(x)} \to 0$ as $\rho \to 0$.

Let us now minimize the bound in (6.4). Two cases need to be considered:

a) There is no constraint on n.

Since (2), (3), and (4) tend to zero with $n \to \infty$ when $\rho < 0$, $s > 0$ and $\psi = \mu'(\sigma) < \psi(\lambda_s)$, n can be chosen large enough to make (2), (3) and (4) negligible with respect to (1). We want, therefore, to minimize (1) under the constraint that (6.5) is

satisfied, i.e.

$$\begin{cases} \min\limits_{\rho, q(\cdot), s} \sum\limits_{e} P(e) W_\rho(x|e) \delta_1(e,x) \\ \text{subject to} \\ \sum\limits_{e,x} P(e) W_\rho(x|e) \log \dfrac{W_\rho(x|e)}{q(x)} = \psi = \mu'(\sigma) < \sum\limits_{xx'} q(x) Q_{\lambda_s}(x'|x) \log \dfrac{Q_{\lambda_s}(x'|x)}{q(x)} \end{cases} \tag{6.6}$$

Now suppose ρ_0, $q_0(\cdot)$ and s_0 are solutions to (6.6), then

$$\sum_{x,x'} q_0(x) Q_{\lambda_{s_0}}(x'|x) \log \frac{Q_{\lambda_{s_0}}(x'|x)}{q_0(x)} \leq \sup_{q(\cdot),x} \sum_{xx'} q(x) Q_{\lambda_s}(x'|x) \log \frac{Q_{\lambda}(x'|x)}{q(x)}$$

$$\therefore \Sigma_{xx'} q_0(x) Q_{\lambda_{s_0}}(x'|x) \log \frac{Q_{\lambda_{s_0}}(x'|x)}{q_0(x)} \leq C$$

with equality only if $q_0(\cdot)$ and s_0 are solution to 5.11. If that is so we will say that the source is matched with the channel. If it is not so,

$$C - \sum_{xx'} q_0(x) Q_{\lambda_{s_0}}(x'|x) \log \frac{Q_{\lambda_{s_0}}(x'|x)}{q_0(x)}$$

could be used as a measure of the mismatch between Source and Channel. $\therefore$ in general $\min_{\rho, q(\cdot), s} \Sigma_e P(e) W_\rho(x|e) \delta_1(e,x) \geq R^{-1}(C)$, by definition of $R(D)$.

THEOREM IV. Let S be a constant, memoryless source which generates messages $e \in E$ with a fixed probability $P(\cdot)$. Let $d(\cdot,\cdot) : E \times A \to \text{Reals}$ be the loss function attached to S and A, the set of feasible actions. Let $(X, p(y|x), Y)$ be a discrete, memoryless channel of capacity C.

Let a vocabulary $u = \{u_1, \cdots, u_M\}$, $M = \underline{e}^{n\psi}$, $u_m = (x_{m1}, \cdots, x_{mn})$ be generated with probability $\text{Prob}\{\ \} = \prod_{m=1}^{M} \prod_{k=1}^{n} q(x_{mn})$. Let the source messages e^n be mapped onto u using the encoding function $\psi_1(\cdot)_u$. Let the outputs y^n of the channel be decoded by $\psi_2(\cdot)_u$. Then:

$$
\begin{aligned}
\underset{u,T(\cdot),e^n}{E} \{d(e^n, \psi_2[T(\psi_1(e^n)_u)]_u)\} \leq & \sum_e P(e) W_\rho(x|e) \delta_1(e,x) \qquad (1) \\
& + d_{\max} [\underline{e}^{-n} + n^{\frac{GH}{2}} \beta(\rho) \underline{e}^{-n[\sigma\psi - \mu(\sigma)]}] \qquad (2) \\
& + 3\{\sum_x q(x) g'(x,s)\} \underline{e}^{-n[\psi(\lambda_s) - \psi]} \qquad (3) \\
& + \delta_2 \max \sum_x q(x) \underline{e}^{-n[sg'(x,s) - g(x,s)]} \qquad (4)
\end{aligned}
\tag{6.7}
$$

where $\rho, q(\cdot), s$ minimize

$$\begin{cases} \sum_{e} P(e)W_{\rho}(x|e)\delta_1(e,x) \\ \text{under the constraints } \rho < 0, \quad s > 0 \quad \text{and} \quad 0 < \sigma < 1 \qquad (6.8) \\ \sum_{e,x} P(e)W_{\rho}(x|e) \log \frac{W_{\rho}(x|e)}{q(x)} = \psi = \mu'(\sigma) < \sum_{xx'} q(x)Q_{\lambda_s}(x'|x) \log \frac{Q_{\lambda_s}(x'|x)}{q(x)} \end{cases}$$

and $\Sigma_e P(e)W_{\rho}(x|e)\delta_1(e,x) \geq R^{-1}(C)$ with equality only if Source and Channel are matched.

COROLLARY III. There exists a code (ψ_1, ψ_2) whose expected loss satisfies (6.7)

b) the block length n is constrained to be at most equal to N.

THEOREM IV'.

$$\bar{d} \leq \min_{\rho, q(\cdot), s, n, \psi} \{ \sum_{\rho} P(e)\gamma'(\rho,e) + d_{max} [\underline{e}^{-n} + n^{\frac{GH}{2}} \beta(\rho) \underline{e}^{-n(\sigma\psi - \mu(\sigma))}] \qquad (6.9)$$
$$+ 3 \{ \sum_{x} q(x)g'(x,s) \underline{e}^{-n[\psi(\lambda_s) - \psi]} + \varepsilon_{2\,max} \sum_{x} q(x) \underline{e}^{-n[sg'(x,s) - g(x,s)]} \}$$

with $\rho < 0$, $s > 0$, $n \leq N$, $\psi = \mu'(\sigma)$, $0 < \sigma < 1$.

The minimization is made easier by the fact that all the function in the bound are well behaved, either increasing or decreasing in the various parameters.

SECTION 7

Treatment of the General Problem with Certain Properties of the Source and the Channel Assumed

In this section, we deal with a more restrictive set of assumptions, namely:

- the source S is binary, memoryless and uniform, i.e., $E = \{1,2\}$, $\{e_k\}_{k=1}^{\infty}$ is a sequence of independent identical random variables and $\Pr\{e_k = 1\} = \Pr\{e_k = 2\} = \frac{1}{2}$.

- $A = \{1,2\}$

- The loss function is symmetric in the following sense:

DEFINITION. A loss function is symmetric if each row of the loss matrix $[d(e,a)]$ contains the same set of numbers $d_1, \cdots, d_H$ and each column of $[d(e,a)]$ contains the same set of numbers $d'_1, \cdots, d'_G$.

$[d(e,a)] =$

e \ a	1	2
1	0	d_{max}
2	d_{max}	0

We will take $d_{max} = 1$ without loss of generality, so that $E_{e,a}\{d(e,a)\} = \text{Prob}\{\text{error}\}$ per message, since $d(e,a) = 1 - \delta_{ea}$, where δ is the kronecker symbol.

- The channel is binary, symmetric and memoryless.

DEFINITION. A channel is symmetric if each row of the channel matrix $[p(y|x)]$ contains the same set of numbers $p'_1, \cdots, p'_J$ and each column of $[p(y|x)]$ contains the same set of numbers $q'_1 \cdots q'_I$.

$$P(y|x) = \begin{array}{c|cc} x \backslash y & 1 & 2 \\ \hline 1 & 1-p & p \\ 2 & p & 1-p \end{array} \qquad p < \frac{1}{2} \qquad \text{w.l.q.}$$

- The utility function is linear in all the criteria, to be given later.

These hypothesis have been selected in such a way that:

- the computations be feasible by hand,
- the number of parameters be reduced to a minimum,
- the results be simple enough to allow a direct reading of the effect of each parameter.

This will enable us to discuss, for this case, the general problem, stated in section 1, of the choice of an optimal Information System (i.e., Encoding, Channel, Decoding), given a Source of messages and the user's utility function.

It is important to note that "binary" can easily be dropped if all the other assumptions are kept.

Let us compute the right hand side of (6.9)

$$\min_{\rho, q(\cdot), s, n, \psi} \Big\{ \underbrace{\sum_e P(e)\gamma'(\rho, e)}_{(1)} + \underbrace{d_{max}\left[\underline{e}^{-n} + n^{\frac{GH}{2}}\beta(\rho)\underline{e}^{-n(\sigma\psi - \mu(\sigma))}\right]}_{(2)}$$

$$+ \underbrace{3\Big\{\sum_x q(x)g'(x,s)\Big\}\underline{e}^{-n[\psi(\lambda_s)-\psi]}}_{(3)}$$

$$+ \underbrace{\delta_2 \max \sum_x q(x)\underline{e}^{-n[sg'(x,s)-g(x,s))]}\Big\}}_{(4)}$$

$$\rho < 0, \; s > 0, \quad \psi = \mu'(\sigma), \quad 0 < \sigma < 1.$$

In this perfectly symmetric situation, the optimal association between X, Y and A is, a priori:

X	Y	A
1	1	1
2	2	2

Likewise the optimal $q(\cdot)$, $\forall$, ρ, s, n, ψ is the uniform distribution $q(x) = \frac{1}{2}$ $\forall$ x.

A. Computation of (1) and (2)

$\delta_1(e,x) =$

e \ x	1	2
1	0	1
2	1	0

$$\gamma(\rho,e) = \log \sum_x q(x)\underline{e}^{\rho\delta_1(e,x)} = \log \left(\tfrac{1}{2}\,\underline{e}^{\rho\times 0} + \tfrac{1}{2}\,\underline{e}^{\rho\times 1}\right)$$

$$\gamma(\rho,e) = \log \tfrac{1}{2}\,(1 + \underline{e}^{\rho}) \quad \forall\, e \tag{7.1}$$

$$\gamma'(\rho,e) = \frac{\underline{e}^{\rho}}{1 + \underline{e}^{\rho}} \quad \forall\, e \tag{7.2}$$

$$\mu(\sigma) = \log \sum_e P(e)\,\underline{e}^{[-\sigma(\gamma(\rho,e)-\rho\gamma'(\rho,e))]}$$

$$\mu(\sigma) = -\sigma\left[\log \tfrac{1}{2}\,(1 + \underline{e}^{\rho}) - \frac{\rho\underline{e}^{\rho}}{1 + \underline{e}^{\rho}}\right] \tag{7.3}$$

$$W_\rho(x|e) = \underline{e}^{\rho\delta_1(e,x)-\gamma(\rho,e)}\, q(x)$$

$$[W_\rho(x|e)] = \begin{array}{c|cc} e \backslash x & 1 & 2 \\ \hline 1 & 1/1 + \underline{e}^{\rho} & \underline{e}^{\rho}/1 + \underline{e}^{\rho} \\ 2 & \underline{e}^{\rho}/1 + \underline{e}^{\rho} & 1/1 + \underline{e}^{\rho} \end{array} \tag{7.4}$$

$$B(\rho,e) = |\rho| + \tfrac{1}{3} + \sum_x \frac{1}{W_\rho(x|e)} = |\rho| + \tfrac{1}{3} + \frac{(1 + \underline{e}^{\rho})^2}{\underline{e}^{\rho}} \quad \forall\, e$$

$$\beta(\rho) = 2\pi\,\underline{e}^{\left[|\rho| + \frac{1}{3} + \frac{(1+\underline{e}^{\rho})^2}{\underline{e}^{\rho}}\right]} \tag{7.5}$$

from (7.1), (7.2), (7.3), (7.5), and (6.9).

$$\therefore (1) + (2) = \frac{e^{\rho}}{1 + e^{\rho}} + \{e^{-n} + n^2 2\pi \, e^{[|\rho| + \frac{1}{3} + \frac{(1+e^{\rho})^2}{e^{\rho}}]} \cdot e^{-n[\sigma\psi + \sigma[\log \frac{1}{2}(1+e^{\rho}) - \frac{\rho e^{\rho}}{1+e^{\rho}}]}} \tag{7.6}$$

By definition

$$\Delta_1 = \frac{e^{\rho}}{1 + e^{\rho}} \tag{7.7}$$

$\sigma = 1$ minimizes (2).

From (4.1.9) and (4.1.10)

$$\log \frac{1}{2}(1 + e^{\rho}) - \frac{\rho e^{\rho}}{1 + e^{\rho}} = - R(\Delta_1) \tag{7.8}$$

We will confine ourselves to $\Delta_1 < \frac{1}{2}$ without loss of generality;

$\therefore$ from (7.7) and (7.8)

$$(1) + (2) = \Delta_1 + \{e^{-n} + 2\pi \, n^2 e^{\frac{1}{3}} \frac{1-\Delta_1}{\Delta_1} \, e^{\frac{1}{\Delta_1(1-\Delta_1)}} e^{-n[\psi - R(\Delta_1)]}\} \tag{7.9}$$

B. Computation of (3) and (4)

$$[\delta_2(x,x')] = \begin{array}{c|cc} x \backslash x' & 1 & 2 \\ \hline 1 & 0 & 1/2 \\ 2 & 1/2 & 0 \end{array} \tag{7.10}$$

$$[\delta_3(x,y)] = \begin{array}{c|cc} x \backslash y & 1 & 2 \\ \hline 1 & 0 & 1/2 \\ 2 & 1/2 & 0 \end{array} \tag{7.11}$$

$$P_s(y|x) = \frac{\underline{e}^{s\delta_3(x,y)}\, p(y|x)}{\Sigma_y\, \underline{e}^{s\delta_3(x,y)}\, p(y|x)}$$

$$P_s(y|x) = \begin{array}{c|cc} x \backslash y & 1 & 2 \\ \hline 1 & \dfrac{1-p}{1-p+p\underline{e}^{s/2}} & \dfrac{p\underline{e}^{s/2}}{1-p+p\underline{e}^{s/2}} \\ 2 & \dfrac{p\underline{e}^{s/2}}{1-p+p\underline{e}^{s/2}} & \dfrac{1}{1-p+p\underline{e}^{s/2}} \end{array} \tag{7.12}$$

$$g'(x,s) = \sum_y P_s(y|x)\delta_3(x,y) = \frac{1}{2}\,\frac{p\underline{e}^{s/2}}{1-p+p\underline{e}^{s/2}} \quad \forall s \tag{7.13}$$

$$\sum_x q(x)g'(x,s) = \frac{1}{2}\,\frac{p\underline{e}^{s/2}}{1-p+p\underline{e}^{s/2}} \tag{7.14}$$

$$E_x\ \{sg'(x,s) - g(x,s)\} = \sum_{x,y} q(x)P_s(y|x)\ \log \frac{P_s(y|x)}{p(y|x)}$$

$$sg'(x,s) - g(x,s) = \log \frac{1}{1 - p + p\underline{e}^{s/2}} + \frac{1}{2} \frac{p_s \underline{e}^{s/2}}{1 - p + p\underline{e}^{s/2}} \quad \forall\, x \tag{7.15}$$

$$Q_\lambda(x'|x) = \frac{e^{\lambda\delta_2(x,x')} \frac{q(x')}{q(x')}}{\Sigma_{x'}\, e^{\lambda\delta_2(x,x')} \frac{q(x')}{q(x')}}$$

$$Q_\lambda(x'|x) = \begin{array}{c|cc} x \backslash x' & 1 & 2 \\ \hline 1 & 1/1 + \underline{e}^{\lambda} & \underline{e}^{\lambda}/1 + \underline{e}^{\lambda} \\ 2 & \underline{e}^{\lambda}/1 + \underline{e}^{\lambda} & 1/1 + \underline{e}^{\lambda} \end{array} \tag{7.16}$$

$$f'(x,\lambda) = \sum_{x'} Q_\lambda(x'|x)\delta_2(x,x') = \frac{1}{2} \frac{\underline{e}^{\lambda}}{1 + \underline{e}^{\lambda}} \quad \forall\, x \tag{7.17}$$

$$\lambda f'(x,\lambda) - f(x,\lambda) = \sum_{x'} Q_\lambda(x'|x) \log \frac{Q_\lambda(x'|x)}{q(x')} \tag{7.18}$$

$$= \log \frac{2}{1 + \underline{e}^{\lambda}} + \frac{\lambda \underline{e}^{\lambda}}{1 + \underline{e}^{\lambda}} \quad \forall\, x$$

λ_s is the maximum λ such that

$$E_{x^n}\{\underline{e}^{-n[\lambda(x^n)f'(x^n,\lambda(x^n))-f(x^n,\lambda(x^n))]}\} \leq \underline{e}^{-nE_{x^n}[\lambda f'(x^n,\lambda)-f(x^n,\lambda)]}$$

$$\left\{\sum_{xx'} q(x')Q_\lambda(x'|x)\delta_2(x,x') \geq \sum_{xy} q(x)P_s(y|x)\delta_3(x,y)\right.$$

But here $\lambda(x^n)$ is independent of x^n, due to the symmetry. Moreover, $\lambda f'(x,\lambda) - f(x,\lambda)$ is independent from x. Therefore :

$$E_{x^n}\{\underline{e}^{-n[\lambda f'(x^n,\lambda)-f(x^n,\lambda)]}\} = \underline{e}^{-nE_{x^n}[\lambda f'(x^n,\lambda)-f(x^n,\lambda)]}$$

$$= \underline{e}^{-nE_x[\lambda f'(x,\lambda)-f(x,\lambda)]}$$

Therefore λ_s is such that:

$$\sum_{xx'} q(x')Q_\lambda(x'|x)\delta_2(x,x') = \sum_{x,y} q(x)P_s(y|x)\delta_3(x,y) \qquad (7.19)$$

i.e.,

$$\frac{\underline{e}^{\lambda_s}}{1+\underline{e}^{\lambda_s}} = \frac{p\underline{e}^{s/2}}{1-p+p\underline{e}^{s/2}} \qquad (7.20)$$

from (7.14) and (7.17), and that relation is true if and only if

$$[Q_{\lambda_s}(x'|x)] = [P_s(y|x)] \tag{7.21}$$

$\therefore$ from (7.15) and (7.18) $\psi(\lambda_s) = sg'(x,s) - g(x,s)$ and (3) + (4) can be written

$$(3) + (4) = \frac{3\underline{e}^{\lambda_s}}{1 + \underline{e}^{\lambda_s}} \underline{e}^{-n[\psi(\lambda_s)-\psi]} + \frac{1}{2} \underline{e}^{-n\psi(\lambda_s)} \tag{7.22}$$

From (7.9) and (7.22) we get

$$\left.\begin{aligned} \bar{d} \leq \Delta_1 + \underline{e}^{-n} + 2\pi\, \underline{e}^{\frac{1}{3}}\, n^2\, \underline{e}^{\frac{1}{\Delta_1(1-\Delta_1)}}\, \underline{e}^{-n[\psi-R(\Delta_1)]} \\ + \frac{3\underline{e}^{\lambda_s}}{1 + \underline{e}^{\lambda_s}} \underline{e}^{-n[\psi(\lambda_s)-\psi]} + \frac{1}{2} \underline{e}^{-n\psi(\lambda_s)} \\ \text{with } 0 < \Delta_1 < \tfrac{1}{2}, \quad s > 0 \\ \psi(\lambda_s) = sg'(x,s) - g(x,s) \end{aligned}\right\} \tag{7.23}$$

Let us denote by $F(\Delta_1, n, \psi, s)$ the upper bound in (7.23). It is easily checked that $F(\cdot)$ is convex $\cup$ in all the parameters.

Now, Δ_1, n, ψ are parameters of encoding and decoding, whereas s is bound to the channel. It is intuitively clear that:

- the encoding gets more difficult as n increases and Δ_1 decreases,
- the decoding gets more difficult as n and ψ increase.

- The channel gets better as the maximum value of s increases.

In this context and short of being able to do better, one might decide to choose an Information System in terms of Δ_1, n, ψ, s and $F(\Delta_1, n, \psi, s)$. One might further decide to attach constant cost coefficients to these parameters and try to maximize a utility function of the form:

$$u(\Delta_1, n, \psi, s, F(\Delta_1, n, \psi, s)) = k_{\Delta_1}\Delta_1 - k_n n - k_\psi \psi - k_s s - k_F F(\Delta_1, n, \psi, s). \tag{6.24}$$

$u(\cdot)$ would have a non-boundary maximum since it is convex $\cap$ in $\Delta, n, \psi,$ and s.

Now, if we assume that the user is not interested in very small n's, i.e., he allows n's large enough so that minimizing (3) + (4) in s would amount to choosing s so as to minimize the exponential terms, that is, he would choose s so as to maximize $\psi(\lambda_s)$. But we have proved in Section 5 that $\max_{q(\cdot),s} \psi(\lambda_s) = C$. We have already maximized $\psi(\lambda_s)$ with respect to $q(\cdot)$. Therefore, for that s_m that maximizes $\psi(\lambda_s)$,

$$\left.\begin{aligned}
&\bar{d} \le \Delta_1 + \underline{e}^{-n} + 2\pi\, \underline{e}^{\frac{1}{3}}\, n^2\, \underline{e}^{\frac{1}{\Delta_1(1-\Delta_1)}\,-n[\psi-R(\Delta_1)]}\, \underline{e} \\
&\qquad + \frac{3\underline{e}^{\lambda_{s_m}}}{1+\underline{e}^{\lambda_{s_m}}}\, \underline{e}^{-n(C-\psi)} + \frac{1}{2}\, \underline{e}^{-nC} \triangleq F^*(\Delta_1, n, \psi, C) \\
&\text{with } 0 < \Delta_1 < \tfrac{1}{2}
\end{aligned}\right\}$$

$$\frac{\underline{e}^{\lambda_{s_m}}}{1+\underline{e}^{\lambda_{s_m}}} = \frac{p\underline{e}^{\frac{s_m}{2}}}{1-p+p\underline{e}^{\frac{s_m}{2}}} : \psi(\lambda_{s_m}) = C.$$

And he would choose Δ_1, n, ψ, C so as to maximize

$$u^*(\Delta_1, n, \psi, C, F^*(\Delta_1, n, \psi, C)) = k_{\Delta_1}\Delta_1 - k_n n - k_\psi \psi - k_C C - k_{F^*}F^*(\Delta_1, n, \psi, C).$$

$F^*(\Delta_1, n, \psi, C)$ gives very explicitly the following asymtotic information:

If $R(\Delta_1) < \psi < C$, there exist vocabularies $u = \{u_1, \cdots, u_M\}$ $M = e^{n\psi}$, such that the average loss due to communication with $(\psi_1(\cdot)_u, \psi_2(\cdot)_u, C)$ be less than $\Delta_1 + \varepsilon(n)$ where $\varepsilon(n) \to 0$ as $n \to \infty$.

BIBLIOGRAPHY

1. Marschak, J., Miyasawa, K., "Economic Comparability of Information Systems", W.M.S.I. working paper No. 85, UCLA, December 1965.
2. Marschak, J., "Problems in Information Economics", W.M.S.I. working paper No. 24, UCLA, November 1962.
3. Marschak, J., "Economics of Inquiring, Communicating, Deciding", W.M.S.I. working paper No. 134, UCLA, January 1968.
4. Marschak, J., "Efficient Choice of Information Services", W.M.S.I. working paper No. 136, UCLA, May 1968.
5. Shannon, C. E., "Coding Theorems for a discrete Source with a Fidelity Criterion," Information and Decision Processes, R. E. Machol, editor, pp. 96-126, McGraw Hill, New York, 1960.
6. Fano, R. N., "Transmission of Information," M.I.T. Press, New York, 1961.
7. Jelinek, F., "Probabilistic Information Theory," Mc Graw-Hill, New York, 1968.
8. Gallager, R. G., "A simple derivation of the Coding Theorem and some applications," I.E.E.E. Trans. on Information Theory, IT-11, 3, Jan. 1965.
9. Ash, R., "Information Theory", John Wiley, New York, 1965.
10. R. J. Pilc, "Coding Theorem for discrete Source-Channel pairs", Ph.D. thesis, Department of Electrical Engineering, M.I.T., February 1967.

11. T. J. Goblick, "Coding for a discrete Information source with a distortion measure", Ph.D. thesis, Department of Electrical Engineering, M.I.T., October 1962.

Security Classification

DOCUMENT CONTROL DATA · R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1 ORIGINATING ACTIVITY (Corporate author) Western Management Science Institute University of California at Los Angeles Los Angeles, California 90024	2a REPORT SECURITY CLASSIFICATION Unclassified 2b GROUP

3 REPORT TITLE

Processing and Transmitting Information, Given a Pay-Off Function

4 DESCRIPTIVE NOTES (Type of report and inclusive dates)

Dissertation--Working Paper

5 AUTHOR(S) (Last name, first name, initial)

Pham-Huu-Tri, Henri Michel

6 REPORT DATE	7a TOTAL NO. OF PAGES	7b NO OF REFS
December, 1968	77	11

8a CONTRACT OR GRANT NO.	9a ORIGINATOR'S REPORT NUMBER(S) Working Paper No. 143
b. PROJECT NO.	
c	9b OTHER REPORT NO(S) (Any other numbers that may be assigned this report)
d	

10 AVAILABILITY/LIMITATION NOTICES Distribution is unlimited:	Western Management Science Institute University of California Los Angeles, California 90024
11 SUPPLEMENTARY NOTES	12 SPONSORING MILITARY ACTIVITY

13 ABSTRACT

An information system is defined as a chain of information services, encoding (processing)...transmitting...decoding (deciding). Each service is a transformer represented, in general, by a stochastic matrix and a cost function. The inputs of "encoding" are the pay-off-relevant events. Actions are the output of decoding, actions and events determine the pay-off. The utility of the services to the user is a function of the pay-off and of the different costs. Efficiently choosing an information system is by definition choosing an information system which maximizes the expected utility.

Communication engineers restricted themselves to information systems with fixed transmitting (channel) and identically zero cost functions. Moreover, they equated the user's utility function with his pay-off function. They handled the problem in the following way:
1. choose first encoding with respect to the source of events and the pay-off function only, 2. choose second encoding and decoding with respect to transmitting only. Encoding is the composition of first and second encoding. However, their approach was inefficient; 1. They neglected the pay-off function in the choice of second encoding and decoding, 2. they arbitrarily broke the original problem into two independent, more accessible, problems.

We also restricted ourselves to information systems with fixed transmitting and zero cost functions and users' utility functions identical to their pay-off functions. But our approach is more efficient because we treated the problem

DD FORM 1 JAN 64 1473 0101-807-6800

Security Classification

DOCUMENT CONTROL FORM

- continued -

of choosing encoding and decoding, given a source of events, a pay-off function and a channel, as a whole. The bounds we obtained should, therefore, be better, at least in all cases where the pay-off function has a wide range of values. We did, however, treat the non-restricted problem with certain properties of the source, the channel and the utility function assumed.